

Decision time at Yucca Mountain

The United States should only proceed with its sole proposed nuclear-waste repository once geological concerns have been addressed and provided that law-makers are convinced that the assessment process has been conducted fairly.

This time it is for real: after years of investigation and prevarication, the US government will soon decide whether it can safely bury spent nuclear fuel and other nuclear waste at a planned repository under Yucca Mountain in Nevada. The decision is the culmination of a vast 20-year exercise to characterize the proposed site and assess its associated risks (see News Feature, page 850).

The nuclear power industry in the United States currently stores most of its spent fuel on the surface, at or near nuclear power stations. The US Department of Energy (DOE) is legally obliged to take the spent fuel and so is under intense pressure to proceed with the Yucca Mountain project — currently the United States' only option for long-term disposal.

The decision on whether to proceed with Yucca Mountain has been a long time coming. Sceptics doubt if 20 years of study at a cost of some \$7 billion has really been necessary to establish the project's viability and safety. But if the outcome of the process is some measure of public acceptance for the project — not least in Nevada — then the wait will have been worthwhile. Indeed, an argument can be made that the government has not yet waited long enough.

The Yucca Mountain project began its tortured existence in an era when public demands for accountability in science-based decision-making were far less onerous than they are today. In those far-off days of the 1970s and early 1980s, the DOE was a cold-war autocracy whose apparatchiks saw little need for consultation either with the general public or with technical advisers from outside its own ranks.

But times have changed since then, and an excellent and timely report (*Disposition of High-Level Waste and Spent Nuclear Fuel: The Continuing Societal and Technical Challenges* — see <http://www.nap.edu>) released earlier this year by the US National Research Council (NRC) describes what the changes mean for nuclear-waste disposal. The report notes that no nation has yet succeeded in building

a permanent store for spent fuel or other highly radioactive waste.

The approach recommended by the NRC's report rests on two pillars: complete openness with the public, and external scientific peer review. The Yucca Mountain process has been strong in the former, at least lately, but weak in the latter. It has been far too reliant on the expertise of the DOE's own laboratories, and not insistent enough on publishing its findings in the scientific literature or on reaching out to experts in the universities and elsewhere. For example, recent work by outside experts on lava flow suggests possible shortcomings in the models that the DOE has used to predict the integrity of the repository in the event of volcanic activity at the mountain.

In the past few years, the project has been gradually evolving in these directions. It has appointed an international panel to review its work, for example — although the panel will not report until after the United States has passed its critical decision point. President George W. Bush is expected to announce his own decision on the matter by the end of the year. It is likely that, armed with a recommendation from the DOE that the site is suitable, he will say the project should proceed. The state of Nevada is then entitled to declare its opposition to the project, and is certain to do so. Finally, under a special law, it will fall to both houses of Congress to vote the project up or down, probably in early 2002.

There is a natural tendency for the technically minded to root for projects such as Yucca Mountain to succeed. Many of those who have been involved in nuclear science and engineering feel that their work is somehow tarnished if a technical problem, such as the safe disposal of spent nuclear fuel, cannot be fixed.

But the paramount considerations must be that the solution is safe and, just as importantly, that the process by which it is derived has been fair. The project should proceed only if the DOE can establish both points. ■

Women worse off (cont.)

The California Institute of Technology needs to address imbalances and frustrations.

With a growing number of women students, and situated in the state that sees itself as America's most progressive, the California Institute of Technology would seem to be an ideal arena to nurture a research community free of gender warfare. But a study reported in this issue (see page 844) shows deep dissatisfaction among the one in ten of Caltech's faculty who are women. The women feel — and the facts suggest — that they have had an anomalously high chance of being spurned when leadership positions were filled; they feel discriminated against by administrators who dole out endowed chairs, which underwrite salaries; and they have now learned that they are paid less for equal work.

The intense atmosphere has led to backbiting. A senior woman faculty's reticence to fight for equal salaries was called "traitorous" by another woman professor. Women faculty who push for their rights were called "young hotheads" by senior faculty. And there is lingering

bitterness among the senior and junior faculty women over battle tactics in the gender war. All this in a state where women have increasingly assumed leadership positions in the corporate and political world.

This past spring, Caltech elected the first woman 'chair' of the faculty. Last winter, the biology division saw its first woman faculty member receive an endowed chair, after some 70% of male biology faculty already had endowments. But the woman biologist's endowment came only after strong lobbying.

As a private university, the onus for improvement rests with Caltech's administration and board of trustees. Its president, David Baltimore, is lauded by a number of women faculty for his leadership and open-door policy of listening to all faculty. Ensuring that women have an equal voice in leadership and committees, bestowing endowments in a way that is demonstrably fair, and giving women equitable salaries would be three major steps forward. ■





Patent clash
Inventor of blue
LEDs sues former
employer
p844



Call of the wild
Siberian project for
domesticating foxes
faces closure
p845



Increased scope
Financial boost
for Australian
sciences
p846



Bared bones
Hints of malaria
found in fifth-
century cemetery
p847

Japan's plans for space merger spark fears for basic research

David Cyranoski, Tokyo

Japan is to merge its three major space laboratories to create a single organization responsible for all space-related research and development.

The merger plan is in keeping with Prime Minister Junichiro Koizumi's vision of streamlining government-funded research (see *Nature* **412**, 364; 2001). But it is already concerning some space scientists, who fear that manned space flight and engineering-based activities will dominate the new organization.

The plan will combine the National Space Development Agency (NASDA) with Japan's main aerospace and space-science laboratories. NASDA is increasingly involved in commercial space applications, such as building and launching communication satellites. The National Aerospace Laboratory of Japan focuses on research areas such as fluid dynamics, whereas the Institute of Space and Astronautical Science (ISAS) conducts space science.

"The merger will bring together divided resources and make it easier for researchers to move freely between projects," says Fujiki Kanji, a senior official at the education ministry, which oversees the three laboratories. Sharing major facilities will make the new body more efficient, adds deputy education minister Takashi Aoyama.

Currently, each laboratory has a different funding basis and administrative standing, which will complicate the merger. Aoyama will chair a committee of laboratory managers and outside experts that will plan the details of the merger before the start of the new fiscal year, next April.

Space commercialization is likely to be a major emphasis for the new organization, Aoyama says. But some researchers worry that basic research could be swallowed up by NASDA programmes such as rocket development and participation in the International Space Station, which account for four-fifths of the combined ¥194 billion (\$US1.6 billion) budget of the three laboratories. "The barriers between our basic research and their more applied research will be broken down, and if money is needed for the larger



Lofty ambitions: Japan hopes the merger will invigorate space projects, including its H-IIA rocket.

projects, they might take it away from ours," says one ISAS researcher.

Others take a more sanguine view, saying that the change could reinvigorate a field that is facing a crisis of confidence after the failed

launch of ISAS' M-V rocket last February, and of two of NASDA's H-II rockets.

► <http://www.isas.ac.jp/e/index.html>

► http://www.nasda.go.jp/index_e.html

► <http://www.nal.go.jp/Welcome-e.html>

Stem-cell list offers sixty-four lines

Jonathan Knight, San Francisco

The US government has published a list of the 64 human embryonic stem-cell lines that can be used by publicly funded researchers under the Bush administration's new rules.

Publication of the list, which the National Institutes of Health (NIH) posted on its website on Monday, ended weeks of head-scratching among researchers over the large number of stem-cell lines that President George W. Bush claimed existed in his 9 August speech. The list names 10 companies and research labs worldwide that have derived cell lines.

But questions remain about the nature and quality of many of the cell lines, and not all the labs are ready to give the cells out.

The highest concentration of stem-cell lines is in Sweden — 19 at Göteborg University and 5 at the Karolinska Institute

in Stockholm. Another 10 cell lines are at two centres in India, with 4 more in Israel.

But CyThera, a San Diego-based company developing stem-cell therapies for degenerative diseases, says it does not know when it will be ready to distribute its nine lines. "We first need to characterize them fully," says its chief operating officer Lutz Giebel.

Audrey Chapman, a science-policy official at the American Association for the Advancement of Science, says the list only begins to address the concerns outlined in an association statement produced on 17 August (see *Nature* **412**, 753–754; 2001).

"The question really is how much more the NIH knows but is not ready to share, or whether they have shared all they know," she says.

► www.nih.gov

Elusive fossil could conceal answer to dinosaur debate

Rex Dalton, San Diego

A Chinese dinosaur fossil of immense potential interest to palaeontologists is circulating between private dealers in Europe, highlighting scientists' concern that the underground trade in such artefacts is inhibiting their study.

The fossil of a psittacosaurid is special because it has outer coverings, known as integuments, protruding along the tail, something not previously observed in the dinosaur group to which it belongs. If it were to be established that these protrusions carried hair or feathers, that could mean that most dinosaurs had these features.

Mark Norell, an expert on Asian dinosaurs at the American Museum of Natural History in New York, says the story is a telling example of how private dealings can hide important scientific discoveries. "This specimen could redefine how we look at dinosaurs," he says. "But we can't say, since no one has studied and published on it. It should be returned to China, where it can be studied in a museum."

According to interviews with palaeontologists and dealers, the fossil has made a mysterious journey through Italy and Germany over the past two years. It was last heard of in the hands of a German dealer.

In recent weeks, the fossil has been the subject of extensive discussion on palaeontology websites, prompted by the publication of a painting thought to show a psittacosaurid in the popular book *Extreme Dinosaurs* by Luis Rey.

There are two orders of dinosaurs: Saurischia and Ornithischia, the latter of which includes psittacosaurids. Some saurischid fossils show integuments, but none have been positively identified on ornithischids, which would make the psittacosaurid specimen an important first.

Last year, the fossil was being kept in a shop called Stone Age in Trieste, Italy. The shop's proprietor, Flavio Bacchia, says that the German dealer, a geologist, had given him the specimen for cleaning. Bacchia says he sent the fossil for study to palaeontologists at Milan's natural history museum.

A museum researcher says that the fossil was examined there in some detail, but that Bacchia then requested its return. Bacchia says he has now returned the fossil to the German dealer, who did not respond to an interview request. Bacchia also says that while the fossil was in Italy, a failed attempt was made to arrange a collaborative study of it with Chinese palaeontologists. ■

♦ www.cmnh.org/fun/dinosaur-archive

Staff survey shows women feel out in the cold at Caltech

Rex Dalton, San Diego

Women staff at the California Institute of Technology are highly dissatisfied with academic conditions there, according to an internal study that has also found they are paid less than men at the prestigious university.

The investigation finds that women are unhappy with the economic support for their positions, their lack of involvement in faculty leadership, and the method by which they are promoted and receive tenure.

"The study shows a lot of dissatisfaction among women," says astronomer Anneila Sargent, chair of the faculty committee that conducted the two-year study. The committee is now preparing recommendations for Caltech administrators about the issues it raises.

Among the 283 academic staff at Caltech during the period studied, only 27 were women. But over one-third of undergraduates and one-quarter of graduate students there are women.

The study found that 54% of women were dissatisfied or had reservations about their career advancement, compared with 27% of men. Almost two-thirds of the women were dissatisfied or had reservations about faculty



Anneila Sargent says Caltech's women staff are unhappy.

committee participation, while less than a quarter of men had reservations and none were dissatisfied.

David Baltimore, the institute's president, says: "Caltech is trying very hard to bring women into parity with men in the academic world. We are very conscious and sensitized to the issues, and will try to address whatever imbalances there are."

Sargent's committee is likely to recommend changes in how women are promoted and given tenure, and in the openness of academic affairs. "When a process is not transparent, there is always room for a perception of injustice," says Sargent.

The study also found that women were paid less than men, although no figures are disclosed in its report. "Clearly, if there are inequities in salaries we will need to take a serious look at that," says Steve Koonin, provost of the institute. ■

LED pioneer seeks reward in court

David Cyranoski, Tokyo

A prominent Japanese materials scientist is suing his former employer for ¥2 billion (US\$16.5 million) over the level of compensation he received for inventing the technology behind blue light-emitting diodes (LEDs).



Out of the blue: Shuji Nakamura's research helped to make blue and green LEDs a reality.

The claim — filed by Shuji Nakamura, now at the University of California, Santa Barbara, against Nichia Corporation of Shikoku — has sent shock waves through Japanese industrial research.

LEDs are semiconductor-based devices that convert electrical current into light. Nakamura is suing Nichia in Tokyo's district court over a patent, filed in 1991, covering a processing method for making the light-producing layer of gallium nitride-based semiconductors. Nichia, where Nakamura worked from 1979 to 1999, used this technology to produce blue and green LEDs as well as blue lasers.

Nakamura is basing his case on Japan's ambiguous patent law, under which researchers should get "fair compensation" for their patented research. Researchers do not usually have contracts with detailed reward provisions. "We need this kind of case to force employers to make clear reward conditions for inventors," says Nakamura.

Nakamura is claiming ownership, in part or in full, of the patent, and wants the technology to be licensed to outside manufacturers. "Nichia will not license it to others," he says. ■

Joint venture on biochips ends in disarray

Quirin Schiermeier, Moscow

A successful scientific collaboration between Russia and the United States for developing microchips to process genetic information has collapsed in acrimony.

The collaboration, which involved the Engelhardt Institute of Molecular Biology (EIMB) in Moscow, the Argonne National Laboratory in Illinois and two US-based commercial partners, had provided substantial funding for the Moscow laboratory in exchange for access to its expertise in DNA chip technology.

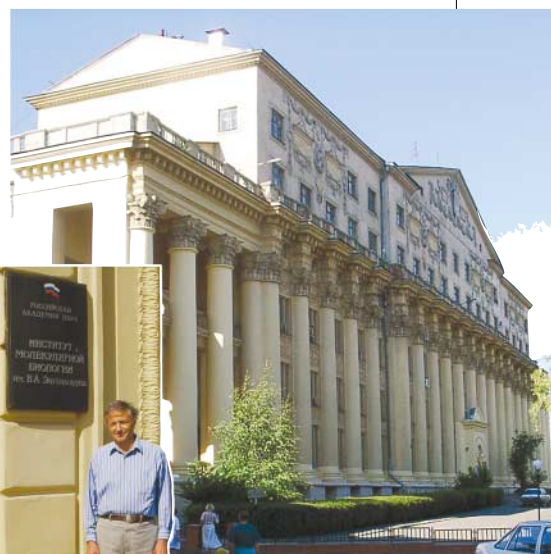
But in recent months, financial support for the Russian laboratory has dried up as EIMB and Argonne argue over the contractual arrangements between the parties. As a result of the dispute, Andrei Mirzabekov has resigned as director of Argonne's Biochip Technology Center. In turn, Argonne has suspended the transfer to the Russian laboratory of some of the royalties it receives from the collaboration's industrial partners.

The collaboration, which began in 1995, has generated almost 40 scientific publications and 30 patents and patent applications. It also triggered a partnership with Packard

BioScience, a Connecticut-based instrument maker, and the electronics company Motorola. The companies have used research from the collaboration to develop commercial microchips based on tiny gels applied on a glass surface which can perform thousands of biological reactions within seconds, and associated instruments to process and analyse results from the chip. Researchers use the technology to detect genetic mutations and for drug design.

Motorola and Packard have contributed almost US\$20 million to the project over five years, making it one of the most successful biotechnology agreements between the US Department of Energy, which runs Argonne, and industry. About 15% of this money went to the EIMB, which also, Mirzabekov says, received funding through a licensing agreement with Argonne.

Argonne officials agree that the arrangement has been a roaring success. "This has been a true model relationship," says one. But according to Bill Ragland, a research manager at Argonne, it is now time to reduce the size of the Moscow research operation supported by the collaboration.



Moving on: director Andrei Mirzabekov (inset) is looking for new partners for Moscow's Engelhardt Institute of Molecular Biology.

Mirzabekov claims that some US officials have turned sour on the arrangement because of security concerns about the possible use of biochip technology in detecting biological warfare agents. Argonne officials deny that this has any connection with their decision.

According to Argonne, royalty payments of about \$500,000 for the year 2000 have been stopped because of contractual disagreements over intellectual property rights on microchip technology. The technology was initially described by a team at the Moscow lab in the 1980s. Argonne rejects Mirzabekov's claim that rights for the technology, which was subsequently developed by the joint venture, should now revert back to the EIMB.

Argonne officials say they are keen to transfer the outstanding money, but first want Mirzabekov to sign a contract assuring them that he will continue to abide by earlier financial agreements. Mirzabekov is unhappy with the contract, saying he already has one, and that Argonne is effectively seeking to renegotiate it. "It makes us very worried about the future, thinking that Argonne could depart from other contracts as well," he says.

Mirzabekov says he has hired lawyers in the United States, but cannot afford to fight a lengthy legal battle with the US lab. He hopes instead that the Russian science ministry will settle the case directly with the US Department of Energy.

Mirzabekov is now searching for new partners in the United States or Western Europe. He hopes to use biochips to detect drug-resistant strains of tuberculosis — a growing health problem in Russia.

It's a dog's life for Siberian foxes

Bryon MacWilliams, Moscow

A unique project in Siberia that has helped researchers understand how animals become domesticated could soon expire from lack of funds.

Researchers at the Institute of Cytology and Genetics in Novosibirsk have compressed into decades a process that might otherwise require thousands of years. Through intense selective breeding, they have turned the silver fox (*Vulpes vulpes*) into an amiable, tail-wagging house pet.

The work, started in 1959 by the late geneticist Dmitry Belyaev, has

demonstrated the close links between behavioural and developmental genetics.

The silver fox — known as the red fox in the United States — is closely related to the dog, but had never been domesticated. The 45,000 foxes so far involved in the Siberian project were chosen by researchers for their propensity to being tamed.

The selection process resulted in the emergence of traits in physiology, morphology and behaviour that are most noticeable in the changes in coat colour and the presence of floppy ears and curled tails.

But the institute lacks the \$10,000 a year it needs to keep the project going. "This is the only kind of experiment of its kind in the world," says Ludmila Trut, head of the research group at the institute. "It would be a tragedy if it were to wither."

The research "is really important because evolutionary change in behaviour is very hard to study," says Deborah Goodwin, deputy director of the Anthrozoology Institute at the University of Southampton in Britain.

The population of breeding foxes has been reduced from 700 to 100 since 1996 and the staff, too, is declining. The project is now overseen by seven researchers of retirement age.

► <http://www.bionet.nsc.ru/ICIG>



Tail of woe: lack of money leaves the project to domesticate foxes facing an uncertain future.

EUGENY V. DREVAL



GEMINI

Bright outlook: Australian astronomers can look forward to more time on Gemini.

Funding bonanza for astronomy and biotech in Australia

Peter Pockley, Sydney

In a bid to answer perennial charges that it has short-changed science, the Australian government is to provide A\$151 million (US\$80 million) over five years for 15 major facility projects.

Astronomers won the largest single grant, gaining A\$23.5 million for studies using the international Gemini telescopes in Chile and Hawaii. But biotechnology was the most favoured field overall, with with total funds of A\$47 million for four projects, reflecting the government's emphasis on research with likely commercial applications.

Intense competition for the grants, which were announced on 21 August, meant that only about one in six of the 86 applicant projects was successful.

Nonetheless, Australian astronomers can now double their observing time on the 8-metre Gemini telescopes. They will also develop instruments to help Australia's bid to host the proposed international Square Kilometre Array radiotelescope.

The second-largest single grant, of A\$18 million, will help to establish a neuroscience facility based primarily in Melbourne for studying and treating schizophrenia and Alzheimer's and Parkinson's diseases. Stem-cell research got a smaller grant for a new cell-engineering centre at Monash University in Melbourne.

Research at synchrotron light sources will continue with a A\$14.8-million grant for Australian beamlines at facilities in Japan and the United States.

Other projects to win backing include facilities for proteomics and genomics, a photonics foundry, a livestock genetic database, a marine station for oil and gas exploration in tropical waters, and a centre for wine research. ■

Canada unveils plans to build nanotechnology centre...

David Spurgeon, Montreal

Canada is set to invest Can\$120 million (US\$78 million) over five years in nanotechnology, creating a new institute based in the western province of Alberta.

On 17 August, Prime Minister Jean Chretien and Ralph Klein, premier of Alberta, unveiled plans for the Institute for Nanotechnology, which will be attached to the University of Alberta in Edmonton. Employing about 200 people, the institute will be funded jointly by the federal and provincial governments.

Federal support for the institute will come from Canada's National Research Council (NRC). It will be the first NRC institute to be governed jointly with the host university. The centre will collaborate closely with the University of Alberta, the NRC says, making joint appointments and sharing research facilities.

Peter Hackett, the NRC's vice-president for research, says he is "overjoyed" by the announcement. "Canada has decided to make a major investment in an emerging field. We did this before as a nation in 1984 in biotechnology — with the Biotechnology Research Institute in Montreal — and that's turned out to be very successful."

Hackett says that the new institute will be "interdisciplinary from day one". The institute will offer "a different culture and a different way of looking at engineering, chemistry,



Thinking small: Peter Hackett says the nanotech institute will improve scientific cooperation.

physics and biology", he says. "Everyone will share this perspective of the nano-world."

Michael Brett, a researcher in nanostructure and thin-film engineering at the University of Alberta, agrees. The cooperative arrangement between the university and the NRC will create "a much richer institute", he says, because of the mix of scientists and graduate students and the availability of university facilities.

The main areas of investigation at the institute will relate to biomaterials, biomedical devices, computing and quantum computing, and catalysts. ■

♦ <http://www.nrc.ca/nanotech>

...as mathematicians beat retreat to Alberta

Erica Klarreich

North American mathematicians will soon be able to immerse themselves in their thoughts at a purpose-built hideaway high in the Canadian Rockies.

When it opens in 2003, the Banff International Research Station will allow invitees to "live, eat and breathe mathematics", says Nassif Ghoussoub, director of the Pacific Institute for the Mathematical Sciences (PIMS) in Vancouver.

Inspired by existing retreats for mathematicians in Oberwolfach, Germany, and Luminy, France, the research station will invite up to 40 mathematicians at a time to participate in intense, five-day workshops.

"We've seen from Oberwolfach and Luminy that this kind of setting is a fantastically successful way to advance a mathematics research agenda," says Philippe Tondeur, director of the mathematics division at the US National Science Foundation (NSF).

The station will cost about \$1.5 million per year to run, and will be supported in roughly equal parts by the NSF, the Natural Sciences and Engineering Research Council of Canada, the Alberta Science and Research Authority and PIMS. The NSF and PIMS have already announced their support, and the other partners are expected to confirm their participation on 24 September.

As many as 1,600 mathematicians could pass

through Banff each year. In addition to the five-day workshops, it will accommodate groups of 10–15 mathematicians to work for up to four weeks on specific research topics. ■

♦ <http://www.pims.math.ca/birs>



H. TURNER/NRC

Green-fingered gang could hold climate key

David Adam, London

Legions of devoted British gardeners may soon be enlisted into climate-change research, as part of an investigation of how weather patterns make their gardens grow.

The UK Climate Impacts Programme (UKCIP), a government-funded organization at Oxford University, is considering asking gardeners to report on the trees, shrubs and flowers in their backyards. The data could help researchers to identify long-term trends in plant development and growth, and then to match them with climate patterns.

The idea reflects rising interest in Britain in phenology — the study of seasonal plant and animal activity. Phenologists use observational data such as the first sightings of migratory birds in spring and first leaf-fall in autumn to identify changes in the biosphere. For example, the average dates of leaf unfolding and colouring have shown that the growing season in Europe is extending (see *Nature* **397**, 659; 1999).

UKCIP officer Megan Gawith says that most research on climate change focuses on the possible effects on commercial agriculture and horticulture. "There has been no research on the domestic garden or heritage garden sector," she says. Together with organizations such as the National Trust and the Royal Horticultural Society, the UKCIP aims to investigate how climate change affects typical features of British gardens.



Research in bloom: there is growing interest in using domestic gardens to study climate change.

Data from domestic gardens may be "very valuable", says Frank Chmielewski, a crop scientist at the Humboldt University of Berlin and coordinator of the International Phenological Gardens, a pan-European network of 62 gardens containing genetically identical trees and shrubs. But he warns that it will take about 10 years to generate data to investigate climate change, and that many observations will be needed, as gardens are rarely typical of the surrounding environment.

"It would be better to advise gardeners to make phenological observations over a larger area," Chmielewski says, "not only in their own garden but also in the gardens of their neighbours." Observers should ignore flow-

ers, he says, because they are more influenced by local conditions than are trees and shrubs.

Britain has around 27 million self-confessed gardeners and a national obsession with the weather, making it ideal for such a study. Yet Mark Schwartz, a phenology researcher at the University of Wisconsin, Milwaukee, cautions that participants will need to be vigilant. "For the data to be useful, the observed events have to be defined quite sharply," he warns.

Gawith says the new scheme will probably be associated with an already established one, such as the UK Phenology Network, which already has 3,000 volunteer observers.

♦ www.phenology.org.uk

Earliest malaria DNA found in Roman baby graveyard

Alison Abbott

Malaria experts are helping archaeologists to work out what caused the deaths of 47 babies whose bodies have been unearthed together in a fifth-century cemetery outside Rome.

The experts think the tiny skeletons — including the bones of 22 miscarried fetuses — may have been victims of *Plasmodium falciparum*, an extremely virulent form of the organism that causes malaria. The Roman site could mark the northern edge of *P. falciparum*'s penetration into Europe.

The archaeologists who excavated the site, led by David Soren of the University of Arizona in Tucson, are intrigued by its unusual configuration. The cemetery, at Lugnano, Umbria, was created in a ruined villa. The bones were found at different depths, with more at higher levels.

"This is not the normal burial pattern for a Roman cemetery," says Soren. There were also signs of witchcraft at the burial site, including puppy skeletons, perhaps meant to ward off demons thought to cause disease. Study of the soil between the

remains indicates that all the babies were buried within a few weeks.

Medics suggest a malaria epidemic as the cause of the deaths. The honeycomb pattern of many of the bones suggests anaemia, a symptom of malaria, and severe malaria is implied in the era's literature.

"Malaria is the most logical explanation for the deaths," says Mario Coluzzi, a parasitologist at the University of Rome.

DNA studies, due to be published next

month in *Ancient Biomolecules*, show the presence of *P. falciparum* DNA in the bones of the oldest infant — the earliest malaria DNA ever identified.

"This is by no means proof that the cause of the infants' death was malaria," says Robert Sallares of the University of Manchester Institute of Science and Technology, UK, who led the DNA study, "but it allows it to remain a contender."

♦ <http://www.coh.arizona.edu/lugnano>



Baby burial: the skeleton of one of the proposed malaria victims unearthed in the Roman cemetery.



Lab life: the British public could soon be better informed about experiments on animals.

Committee calls for end to secrecy in animal research

London The public should have greater access to information about the use of animals in research, a committee advising the British government says.

A report from the Animal Procedures Committee, an independent group that includes practising researchers and animal-welfare representatives, says that a change in the law is needed to address public concerns about secrecy surrounding the licensing and conduct of animal procedures.

The group called for more publication of negative results to prevent repetition and wider dissemination of details of infringements. The committee also says that a clear and concise summary of procedures should be included on licence applications and that these should be made public. The government says that it broadly accepts the recommendations and will act on them.

♦ <http://www.apc.gov.uk>

Johns Hopkins faces fresh probe over study safety

Washington Johns Hopkins University in Baltimore, Maryland, is set to come under further scrutiny after the US Office for Human Research Protections announced plans to investigate a study that was overseen by the university's researchers.

The research into reducing exposure to lead in the home was carried out by the Kennedy Krieger Institute, a paediatric research centre based in Baltimore, but was supervised by scientists from Johns Hopkins. Two lawsuits brought by families involved in the study are currently under consideration in Maryland courts. They centre on allegations that participants in the experiments were not informed of the health risks of the housing in which they were living.

The announcement of the investigation comes a month after the office suspended all federally funded human studies at the university for five days, after the death in June of a healthy volunteer in an asthma study.

Brazil plans to make generic AIDS drugs

London The Brazilian government has announced plans to start producing a generic version of a drug patented by Swiss pharmaceutical company Roche.

The government claims that the country will save US\$35 million a year by producing its own version of nelfinavir, one of a cocktail of drugs used to treat AIDS. It says that negotiations with Roche over possible delivery of the drug at a reduced price had failed to come to a satisfactory conclusion.

Meanwhile, the South African Treatment Access Campaign (TAC) last week took the country's health minister, Manto Tshabalala-Msimang, and nine provincial health ministers to court over their AIDS policies. The campaign demands that nevirapine, which is used to prevent mother-to-child transmission of HIV, be made freely available to pregnant women infected with the virus.

No charge for Japan's cloned mice

Tokyo The mouse clones that formed the backbone of the international effort to make sense of the mouse genome (see *Nature* 409, 685–690; 2001) are to be made available to researchers free of charge.

Dnaform, a spin-off from Japan's Institute of Physical and Chemical Research (RIKEN), based in Tsukuba, near Tokyo, began supplying the 21,000 clones on request on 27 August. Each clone can be used to study an individual piece of complementary DNA (cDNA) and the gene associated with it.

The announcement was made by Yoshihide Hayashizaki, chief researcher on the mouse cDNA project, at last week's Mouse Molecular Genetics Meeting in Heidelberg, Germany. Previously, only researchers collaborating with RIKEN had been allowed to use the clones.

♦ <http://www.dnaform.co.jp>

Canada chews on transgenic food advice

Montreal Canada needs a centralized information system to advise people about the content of food containing genetically modified ingredients, a report from the Canadian Biotechnology Advisory Committee has concluded.

The committee, an independent body set up to advise the federal government on policy issues, also recommended that a chief safety officer be appointed to oversee the safety of genetically modified food and that federal government make a clear distinction between its roles in promoting and regulating biotechnology.

But the committee's decision to call for voluntary rather than compulsory labelling

has angered food-safety groups, such as the non-profit Council of Canadians. Several such organizations boycotted public consultations relating to the report, claiming that the committee was biased in favour of biotechnology interests.

Solar scheme sails on despite missile crash

Washington A private astronomy group has announced plans to press ahead with a scheme to put a solar sail into orbit, despite the failure of a test launch last month.

The Planetary Society, based in Pasadena, California, says that the first of two solar-sail launches will take place early next year. A suborbital launch on 20 July to test the deployment of two 15-metre triangular sails ended when the sail failed to separate from the converted Soviet intercontinental ballistic missile that was carrying it. The rocket and craft crashed on the Kamchatka Peninsula in eastern Russia.

"We would find it a diversion to build another suborbital test craft," said the society's executive director, Louis Friedman, adding that insurance money from the failed test would help to pay for two new rockets and satellites.

♦ <http://www.planetary.org>

French receive taste of crop vandalism

Paris French activists attacked an experimental trial of genetically modified (GM) crops last week, the first significant act of vandalism against GM crops in France.

Around 150 people uprooted insect- and herbicide-resistant plants belonging to US firm Monsanto at a 1,000-square-metre plot in Beaucaire, southern France. The action followed a threat in July by the Confédération Paysanne, the farmers' union led by José Bové, to destroy all open field trials of GM crops. Bové argues that outdoor trials could allow GM crops to spread beyond the trial sites, but says he does not object to tests in greenhouses.

French research minister Roger-Gérard Schwartzberg described the destruction as "unacceptable".



Letting rip: protesters attack Monsanto's transgenic crops in southern France.

Out of sight, out of mind?

Yucca Mountain in Nevada seems like a reasonable place to bury the United State's nuclear waste — but only if the volcanoes there remain dormant. Arguments on that point are about to erupt, reports Colin Macilwain.



A rock and a hard place: Yucca Mountain, proposed site of the US nuclear waste repository.

At the summit of Yucca Mountain, a gentle summer breeze breaks the scorching heat of the Nevada Desert. "People have two reactions when they come up here," says William Boyle, a geologist and senior technical adviser to the US Department of Energy's Yucca Mountain Project, which plans to store nuclear waste at the site. "Some of them say: 'Well, this is pretty well as near to the end of the world as you can get.' The others say: 'Are those volcanoes over there?'"



William Boyle, senior adviser on the Yucca project.

Across the Solitario Canyon, three rusty volcanic plugs protrude from the sandy-coloured rocks that dominate the surrounding landscape. One of them, about 20 kilometres away at Crater Flat, indicates relatively recent volcanic activity in geological terms: it is only about 75,000 years old.

The volcanoes have not been discussed that much during the 20-year characterization to establish whether Yucca Mountain is a suitable nuclear-waste repository. But they will be — and not merely by environmentalists opposed to the project.

The Nuclear Regulatory Commission (NRC) is the government-funded watchdog that must license the proposed facility, and it thinks that the Department of Energy (DOE), which is responsible for getting the

repository built, is seriously underestimating the risk presented by the volcanoes. The argument will come to a head this autumn, just as President George W. Bush is expected to issue his recommendation on whether the project should proceed.

For now, the DOE does not need a firm design proposal for the repository. Rather, it is publishing several documents, including a science and engineering report summarizing all of the studies undertaken during the site-assessment process, that will inform Bush's decision on whether or not to proceed. If he decides to do so, the project team has another two years to finalize its design and submit a licence application to the NRC.

For the NRC to approve the facility, the DOE must meet the standards published earlier this year by the Environmental Protection Agency (EPA). These demand that for 10,000 years after the repository opens, no members of the public should be exposed to more than 15 millirem (0.15 millisieverts) of radiation per year.

Desert manoeuvres

To prepare for its moment of truth, the DOE has mounted one of the largest scientific risk-assessment exercises in history, at a cost of about \$7 billion. Over two decades, its characterization of the repository plan has roped in experts including geologists, microbiologists, seismologists, volcanologists, statisticians, mathematicians, sociologists, politicians, theologians and members of the public. "So far as I know," says Robert

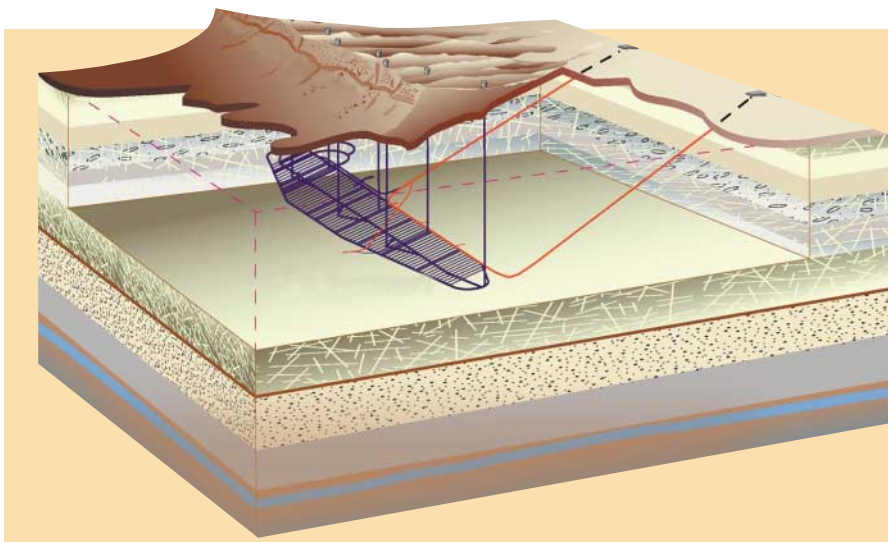
Levich, a geologist and project spokesman, "this has been studied more thoroughly than any other geological site in the world."

The future of US nuclear power hinges on the process. If there is no repository and high-level radioactive waste has to be stored locally on the surface, as it is now, it is unlikely that the United States will build any more nuclear power stations.

Given such high stakes, attempts to find a home for the nuclear waste got off to a rocky start. In the early 1980s, a process was under way to select a suitable repository from a number of reluctant contenders, including salt beds in Texas or Louisiana, basalt in Washington state and granite in the north-eastern United States. But political interventions excluded the other options, and in 1987 Congress decreed that only the Nevada site would be considered.

Many experts say that Yucca Mountain was emerging as the most technically suitable site in any case. "Yucca Mountain was the best site at the time, because of the lack of water," says Michael Voegele, a geological engineer who has worked on repository designs for more than 20 years. He is now chief scientific officer for Bechtel SAIC, the Las Vegas-based private contractor that helps to manage the site-characterization project for the DOE.

The DOE's plan is for Yucca Mountain to accommodate high-level radioactive waste about 300 metres below the desert floor, and yet 300 metres above the water table. It will be housed inside stable, reconsolidated volcanic ash — or tuff — which has lain more or



Mountain hideaway: the nuclear waste will sit 300 metres below the desert in Yucca Mountain.

less undisturbed for several million years.

Inside the experimental tunnel that the DOE has bored some 8 kilometres into the mountain, the walls are dry and pock-marked with thousands of metal plaques that mark holes from which geological samples have been extracted. But the appearance of dryness is deceptive. It is the tunnel ventilation that keeps the surface of the walls dry; the rock behind contains about 10% water by volume. This slowly seeps down from the sporadic rain — 15 centimetres a year, on average — that lands on the desert above.

Tunnel vision

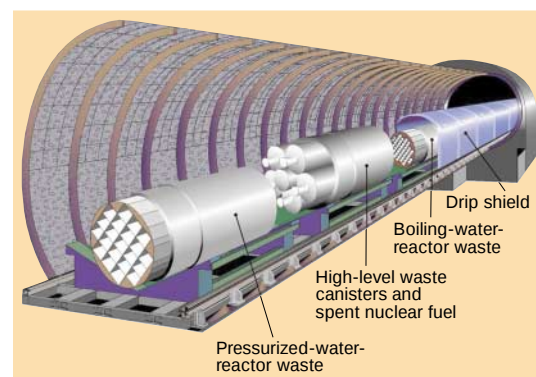
Concerns about the water and doubts about the wisdom of burying the waste irretrievably have caused the DOE to abandon its original plan for the repository, which would have involved filling tunnels in the rock with nuclear waste and permanently sealing them up. Instead, every disused fuel rod and shaft of glass into which waste has been sealed will be housed in multilayered metal canisters, weighing 20–70 tonnes when full, which will sit on the tunnel's floor (see diagrams, above).

The effectiveness of the canisters is being probed by members of the Nuclear Waste Technical Review Board (NWTRB), an official watchdog set up by Congress to keep a close eye on the project. Under President Bill Clinton, the board gradually filled up with members who were sceptical of the project's early self-assurance. "The NWTRB has had a major impact on the project," says Kevin Crowley, staff director for radioactive-waste management at the National Academy of Sciences. "They've pushed for fuller consideration of uncertainties."

At a meeting in Las Vegas in June, for example, NWTRB members questioned the integrity of the DOE's models of how the canisters will behave. Each of the \$1-million cans features a protective layer made of a heat-resistant nickel alloy called alloy 22. Like alu-

minium, this protects itself from corrosion by accumulating a thin, passive oxide layer on its surface. By extrapolating corrosion rates recorded in laboratory tests over two or three months, DOE scientists believe they know how the canisters will fare over 10,000 years. But Paul Craig, a retired engineering professor formerly at the University of California, Davis, who serves on the NWTRB, is not convinced. "To extrapolate four or five orders of magnitude without some kind of mechanism is just not credible," he says.

Discussions between the NWTRB and the DOE have also focused on the extent to which heat emitted by the waste will keep the storage tunnels dry. Last year, the NWTRB asked the project to consider the comparative merits of a radically different approach — a low-temperature repository, in which the fuel canisters would not be crammed so tightly in the tunnels and would be air-cooled during the first few decades of storage, reducing the rate at which they would corrode. Uncertainty about the percolation



Tubular cells: a range of canisters holding radioactive waste will line the floors of dozens of tunnels bored through the mountain.

of water through the tuff has also led project planners to consider a \$4-billion titanium 'umbrella', or drip shield, to protect the canisters from direct exposure to dripping water.

DOE officials remain confident that these uncertainties will not throw the Yucca Mountain Project off course. Even taking the most pessimistic views about the ingress of water and the integrity of the canisters, the mean annual dose that could leak from the repository during the 10,000 years still falls way below the upper limit of 15 millirems. Residents of the community whose water supply would be most affected — Lathrop Wells, a truckers' stopover mainly comprising two filling stations and a brothel — would be exposed to only one-millionth of this dose.

But the risks posed by volcanic activity may be a different matter. So far, these have not drawn much attention from the NWTRB, although it will examine them at a special meeting in Las Vegas on 12 September. Although the prospect of the repository being breached by lava is remote, the consequences could be catastrophic.

The DOE contends, based on studies in the 1980s by Bruce Crowe at the Los Alamos



The hole picture: an 8-kilometre tunnel has been bored through Yucca Mountain to allow for tests.

US DOE

► National Laboratory in New Mexico and others, that the chance of such a breach occurring is just over one in a hundred million per year — a sufficiently small risk that some scientists would be happy to discount it. But the NRC puts the risk at one in ten million per year, which researchers at the NRC say cannot be ignored.

The disagreement stems from the methods used to estimate the risk of an eruption occurring at different areas around Yucca Mountain. The original DOE assessment assumed that all areas had an equal probability of experiencing a volcanic eruption, says John Trapp, an NRC geologist. Trapp and his colleagues at the NRC's Center for Nuclear Waste Regulatory Analyses (CNWRA) at the Southwest Research Institute in San Antonio, Texas, used a more complex model that takes into account the fault characteristics of the rock in different areas (C. B. Connor *et al.* *J. Geophys. Res.* **105**, 417–432; 2000).

The NRC researchers also believe that the DOE is severely underestimating the fallout that would result from volcanic activity. At the heart of this issue is the way in which an eruption would damage the repository.

Dyke dynamics

Eruptions are caused by a vertical wall of lava, called a dyke, that rises towards the surface until it breaks through at the weakest point. At present, the DOE models assume that a dyke would cut cleanly through the repository, expelling the contents of a few waste canisters into the atmosphere. The models anticipate that the radioactive fallout from such an eruption would scatter onto the surface and eventually arrive in the water supply. They predict an average dose at Lathrop Wells of up to a millirem per year — far higher than the projected dose from any other cause, but still an order of magnitude less than the limit permitted by the EPA rule.

But according to Trapp and other volcanologists, a dyke rising through the repository would move at only about a metre per second — and so would be liable to flow through many of the tunnels, heating their contents and perhaps releasing much larger quantities of radioactive waste. “They are equating a dyke to an eruption,” says Trapp. “But a dyke goes real slow.”

Trapp's assessment appears to be backed up by two unpublished papers by groups led by Andrew Woods at the University of Cambridge, UK, who studies lava flow. In a paper currently under review at the *Journal of Geophysical Research*, Woods and Onno Bokhove, a mathematician at Bristol University, model the behaviour of a dyke rising through the repository. They conclude that the shock waves created by the dyke would be powerful enough to break open the end of some repository tunnels and allow most of the contents to be expelled to the surface.



Digging that scene: project engineers watch the completion of the main test tunnel at Yucca Mountain.

Although the paper finds that most of the canisters would not be breached directly by the shock wave, a second paper by Woods and his collaborators at CNWRA, which has been submitted for publication in *Eos*, finds that after the initial shock wave has passed, most canisters would fail as a result of being heated — and weakened — by the lava flow.

The combined implication is that far more waste would be expelled into the atmosphere by an eruption than the DOE has allowed for. NRC scientists have not yet published an estimate of how many canisters of waste may be released in this way, but Brittain Hill, a volcanologist at CNWRA and co-author of the paper submitted to *Eos*, notes that there will be about 150 canisters in each parallel tunnel at the depository, and that a dyke would probably cut into about 10 tunnels. “The DOE models have a dose of 1 millirem from 10 waste canisters,” he says. So for this scenario, the potential fallout could clearly exceed the EPA's limit.

Grounds for appeal

Little headway was made when NRC and DOE volcanologists met in a Las Vegas hotel on 22 June to discuss the implications of the unpublished work. Greg Valentine, a geologist at the Los Alamos National Laboratory who works on the DOE project, says that the NRC's points “are credible, to the extent that we can't just dismiss it. It is something that needs to be looked at.” But Trapp and others argue that the DOE scientists have not fully taken the issues on board. “The meeting was a tremendous disappointment,” he says. Hill claims that the DOE is “giving short shrift to what they admit is the only source of risk.”

The DOE is having to get to grips with the issue just as Bush prepares his decision on whether the repository should proceed. Senior Republican politicians support the project, and Bush is widely expected to

approve it. The Nevada state government then has the right to veto the plan, which it is certain to exercise. But if such a veto is issued, the final say on whether the project can proceed will rest with the US Congress.

Both houses would need to vote for the project by a simple majority, and recent voting history suggests that they are prepared to do so. But since the Democrats wrested back control of the Senate in May, Senator Harry Reid (Nevada, Democrat) has threatened to use his new position as Senate Majority Whip to try to block the project.

In such circumstances, project advocates can ill afford the arrival of a credible scientific challenge to their assumptions about the site's safety. And with just such a challenge looming, some observers are surprised that the DOE has not done more to revisit its models of volcanic activity. As the National Academy of Science's Crowley puts it: “When your regulator says you've got a problem — you've got a problem.” ■

Colin Macilwain is *Nature's* news editor.

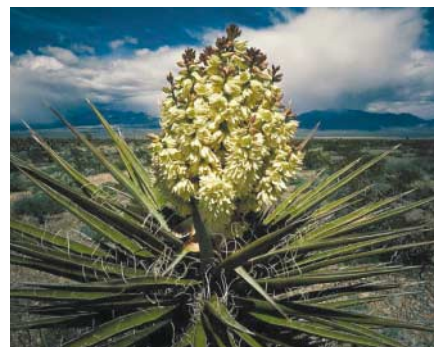
Yucca Mountain Project ♦ <http://www.ympp.gov>

Nuclear Regulatory Commission High-Level Waste Program

♦ <http://www.nrc.gov/NMSS/DWM/hlw.htm>

Nuclear Waste Technical Review Board

♦ <http://www.nwtrb.gov>



A yucca plant in the Nevada Desert, from which the mountain and the project gets their names.

Brought down to Earth

With its unique access to Mir cosmonauts, Moscow's Institute for Biomedical Problems was a world leader for space biology. But now it is working under greatly diminished circumstances, says Quirin Schiermeier.

When the flaming remnants of the Mir space station plunged into the South Pacific on 23 March it marked not only the end of a successful mission, but of an era of human space flight. This expensive and risky pursuit, no longer the ambition of single nations, is now a global partnership.

For the next decade, human activities in space will be dominated by the International Space Station (ISS). And Anatoli Grigoriev, director of the Institute for Biomedical Problems (IBMP) in Moscow, has mixed feelings about this state of affairs. Without the ISS, his institute would have no future. Yet the IBMP's involvement in the project will be very different from the scientific autonomy it enjoyed on Mir. "We used to be hosts, but now we are only guests," Grigoriev laments.

Grigoriev's career has spanned the rise and fall of the Russian space programme. He joined the IBMP as a medical graduate in 1966, just five years after Yuri Gagarin's pioneering flight had electrified the nation. Working on the regulation of blood salt levels under microgravity, he rose through the ranks until he was appointed to his present position in 1988.

The IBMP, with a complement of some 4,000 scientists, engineers and technicians, was then at its zenith. Although the cracks were beginning to show, the Soviet Union was still a genuine superpower. The space programme, for which the IBMP provided full medical support, was the nation's pride and joy, and the first module of Mir had been launched just two years before. "That time was a unique and fulfilling experience," says Valeri Bogomolov, one of the institute's deputy directors.

Through experiments on animals sent



Run down: the Institute for Biomedical Problems in Moscow is a shadow of its former self. Run by Anatoli Grigoriev (below), it has housed well-known scientists, including Inesa Kozlovskaya (bottom).

into orbit, the IBMP had built up its expertise on the biological effects of space flight. But it was the institute's work on Mir cosmonauts, investigating the physiological effects of long-duration space flight, that dominated its agenda under Grigoriev's leadership. IBMP scientists studied how the body adapts to microgravity, and developed countermeasures, such as exercise regimes and gravity-simulation suits. Many other aspects of space research — including water conservation, microbiology and radiation biology — were also tackled.

Decline and fall

But this work was set against a background of steady decline. As research budgets dwindled after the break up of the Soviet Union, Grigoriev's institute had farther to fall than most. Today the IBMP, housed in a series of run-down brick buildings on the outskirts of Moscow, has lost 90% of the staff it had in 1988.

The IBMP's fortunes were mirrored, more publicly, by those of Mir. The station was originally intended to operate until 1991. But when it became clear that Russia could not afford to launch a successor, the space programme's efforts became focused on keeping Mir alive. More modules were added, the last in 1996. In the meantime, Mir cosmonauts set endurance records: Valeri Polyakov, now a deputy-director at the IBMP, stayed in orbit for an unmatched 437 days and 18 hours in 1994–95, conducting medical observations on fellow crew members.

But by 1997 the rot had set in. Mir suffered a fire, computer failures and a collision with a supply vehicle that depressurized one module. In the Western media, Mir was lampooned as an orbiting rust bucket. Its



commander, Vasili Tsibliev, whose health declined as disaster followed disaster, was presented as a pathetic figure, in contrast to the heroic portrayal of the NASA astronaut Michael Foale, who was on Mir at the same time. For staff at the IBMP, it was a dispiriting experience. "It was a very difficult situation," says Bogomolov.

The station's demise has deepened the depression at the institute. The IBMP is now preparing for research on the ISS and, in principle, Russia is an equal partner. But in practice, financial constraints have made



Former glories: Russian–American crews on board the Mir space station. On the right, commander Vasili Tsibliev (right) is shown with Alexander Lazutkin (centre) and American Michael Foale.



the former superpower a space wallflower.

Russia's Zvezda service module, docked onto the ISS in July 2000, could be financed only by selling 4,000 hours of Russia's total research time aboard the first phase of the ISS — more than half of its total allocation.

Many of the IBMP's staff fear that this deal, struck in 1998, will be the start of a larger sell-off. Russia, although invited, is not participating in the International Life Science Working Group that is managing the selection of biology and medical research proposals for the ISS. Russia says it cannot afford to join this club, whose members have pooled their experimentation time and facilities. Staying away gives Russia the option of selling more of its allocated research time. But it also means that Russia will be scientifically isolated on the ISS.

Isolation station

That is unfortunate, because the IBMP's separation from the international scientific community has been a long-standing problem. The institute does have some researchers who are recognized as world leaders in their fields — neurophysiologist Inesa Kozlovskaya, for example, has published more than 250 papers in English-language journals. But she is an exception. Founded as a military institute, much of the IBMP's research was published only in Russian, and often restricted to reports delivered to senior officials. Tellingly, when *Nature* handed Grigoriev the latest issue of the journal, it was the first copy he had ever opened.

Since the end of the Cold War, international collaborations have opened up. In 1991, NASA, the US National Space Biomedical Research Institute in Houston and the IBMP agreed to exchange physiological data. Many Western space agencies sent astronauts to Mir, and ground-based collaborations have also been established (see 'Cosmonauts behaving badly', right).

Most scientists who have collaborated with the IBMP are positive about the experience. But they warn that good personal contacts are needed to ensure that things run smoothly. "It is essential that the modes of operation of a joint project, its scientific

goals, and the methods and statistical tools to be applied are defined in detail," says Rupert Gerzer, director of the Institute of Aerospace Medicine in Cologne, part of DLR, the German aerospace research agency. "Otherwise you can experience unpleasant surprises."

Ary Goldberger of Harvard Medical School has carried out research on heartbeat dynamics using IBMP data. "The Russian data were invaluable," he says. "Immortalizing their unique data in an open archive should be a major priority." The Russian data are not always well collated, says Goldberger. But he adds that he has also experienced problems in obtaining data from NASA.

Grigoriev works hard to dispel reservations about working with the IBMP, and points to the institute's unique experi-

ence in providing medical support for long-duration space missions. "Much is hard for us to do these days," he says. "But without Mir there would be no ISS."

"Russia's long-term background in the area is very valuable," says Victor Schneider, of NASA's life-sciences division. "But the Russian programme, limited in scale, has only whetted our appetite."

Grigoriev agrees that additional research is necessary, in particular on bones, red blood cells, metabolism, and the cardiovascular and immune systems. How much of a contribution the IBMP's scientists make to this work may depend on the health of Russia's fragile economy. But Grigoriev believes that western scientists and space agencies have more to gain from working with the IBMP than just access to its data from Mir. "There are differences in scientific traditions, customs and ambitions," he says. "We have not yet learned the best from each other."

Until now, cultural differences in the approach to human space-flight have tended to cause tensions, with NASA's buttoned-down approach clashing with the Russian penchant for improvisation. But as long as the IBMP remains part of the ISS, Russia's cosmonauts will have the chance to show that they still have something to contribute. ■

Quirin Schiermeier is *Nature's* German correspondent.

♦ <http://www.ibmp.rssi.ru/webpages/engl/welcom.html>

Cosmonauts behaving badly

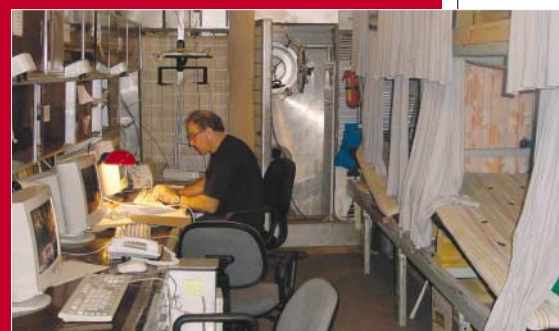
It is 2016. A highly trained multinational crew is midway through an eight-month flight to Mars. But tensions are rising. Russian and American crew members start yelling at each other, then a fight breaks out. Mission control is helpless.

Such scenarios cannot be ignored, suggests a space-flight simulation experiment carried out between June 1999 and April 2000. Three test groups spent prolonged periods in a mock-up of the Mir space station at the Institute for Biomedical Problems (IBMP). One group of four Russians was cooped up for 240 days; two other international groups joined them for 110 days each. The crew members' behaviour, health and interactions were recorded and analysed by teams of Russian, Canadian, German and Japanese psychologists.

Conflicts occurred between different crews and individual crew members. Tensions culminated at a New Year's party in the camera-free kitchen. After a fist fight between two members of the Russian crew, one of the combatants tried to embrace and kiss a female Canadian member from the other group — the only woman on board — against her will.

Following the incident, Russian mission control decided to close a hatch to separate the two groups, and a Japanese crew member quit the experiment. When the Canadian eventually emerged from the Mir mock up, she told the press that she had been sexually harassed, and is now suing the Canadian Space Agency, the IBMP and Vadim Gushin, the study's scientific coordinator.

The New Year incidents also generated distrust among the IBMP's collaborators. Gushin accepts that there were communication problems, but prefers to stress the lessons to be learned for the International Space Station. "Lack of confidence in the Russian mission control caused a constant feeling of psychological discomfort among non-Russian crew members," he says. Rather than the current practice, in which NASA and the Russian Space Agency operate separate mission-control teams, Gushin argues that the teams should be integrated to include representatives of all nationalities on board.



Tensions rose during tests on this mock-up of Mir.

Collaboration with Japan could be more tempting

A pilot scheme for postdocs, blending mobility with security, ought to be expanded.

Sir— Your News story “Japan aims to forge stronger European links” (*Nature* **411**, 875; 2001) clearly identifies why it is difficult to recruit foreign scientists to work in Japan.

The Japan Society for the Promotion of Science (JSPS) has been running a pilot ‘2 + 2 programme’, which allows a British researcher to spend two years in Japan with the assurance of two years’ subsequent funding in the United Kingdom. To promote greater collaboration between research in the European Union (EU) and Japan, this should now be established in other European countries. I would not have applied for this scheme without the return element, yet have already felt the benefits.

Researchers tempted by a period of research in Japan, or wishing to establish a long-term collaboration, face other problems in addition to those mentioned in your News story. Permanent university

staff have heavy teaching loads which, combined with family commitments, allow only the fortunate to take long-term sabbaticals. Postdoctoral researchers are far more mobile (often not from choice). Schemes such as the 2 + 2 programme can turn this mobility into a positive factor, helping them establish long-term collaborations early in their careers, while providing a degree of security.

Unfortunately, much may depend upon the Japanese economy. Provision under the JSPS scheme for Japanese language lessons has recently been reduced, and the tax-free salary status is under threat. Several other improvements could make a significant difference to the quality and number of applicants, for example one paid-for trip home during the two-year period, and greater opportunity for overseas research collaboration. Changes are also required in the host countries — I am unable

to contribute to my university pension while in Japan, for example.

Other schemes for EU–Japanese collaboration should be open to all, unlike Britain’s otherwise excellent Biotechnology and Biological Sciences Research Council (BBSRC) Japan Partnering Awards (<http://www.bbsrc.ac.uk/international/bbsrc/japan.html>), which restricts applicants to current BBSRC grant holders or employees.

Although you report that in the past three years only two people have applied for JSPS funding from the EU, the JSPS does have other types of grant allotted for specific countries or institutions within Europe, and it expects to fund 440 fellowships for the financial year 2001 (see <http://www.jsp.go.jp/e-home.htm>).

Angus Davison

Department of Ecology and Evolutionary Biology, Biological Institute, Tohoku University, Aramaki-Aza-Aoba, Aoba-ku, Sendai 980 8578, Japan

Could a website teach communication skills?

Sir— The Opinion article “Learning to speak and write” (*Nature* **411**, 1; 2001) points out that scientists need to be taught how to communicate the results of their research effectively; in Correspondence, D. A. Watson and his colleagues and M. Attrill (*Nature* **411**, 992; 2001) refer to courses where this is being done.

Young scientists who do not have English as their first language are keenly aware of a further obstacle: the need to acquire these skills in English. Without them, many are media-shy and feel handicapped in effectively communicating the full significance and scope of their work, even if they have good science-communication skills in their native languages.

The Internet is changing the pace and practice of disseminating scientific information (*Nature* **410**, 1023–1025; 2001), but the basic skills of scientific writing are still indispensable. To induce more young scientists to improve their scientific writing skills, the success story of the team running the web-based *Journal of Young Investigators* as discussed in your News Feature (*Nature* **411**, 13–14; 2001) must be repeated by others, launching similar affordable Internet journals in other languages. This will go a long way in building the younger generation’s confidence and skills.

In the absence of any formal training in

scientific writing, it takes weeks or more for scientists to prepare their first few manuscripts in English. Publication will often be further delayed when journal editors ask for manuscripts to be rewritten. In an era when speed of communication is vital, the global scientific community needs a freely accessible website that will help people learn science-communication skills through home study, and assist them in evaluating the level of proficiency they are acquiring.

Virendra K. Bhasin

A. R. L. Teixeira’s Chagas Disease Multidisciplinary Research Laboratory, Faculty of Medicine, University of Brasilia, Brasilia-DF, Brazil

Curious effects created by reversal of colour

Sir— The inadvertent reversal of Masaccio’s *Trinity* (see above left for correct orientation) in my essay “Maintaining Masaccio” (*Nature* **412**, 382; 2001) made nonsense of my point about the relationship between the asymmetrical composition and the off-centre approach by the spectator, but the error may yet be turned to good account.

As Leonardo da Vinci recognized, reversing a composition can make it look very different and even ‘wrong’: “When you are painting, you ought to have by you a flat mirror in which you should often look at your work ... [I]t will seem to be by



Shades of meaning: does the placing of colours give a clue that left is right and right is wrong?

the hand of another master and thereby you will better judge its faults.”

This property of a reversed image has been widely recognized, with respect to the arrangement of forms, but it occurs to me, looking at the flipped *Trinity*, that colour might also be involved. Masaccio’s carefully orchestrated asymmetries of red and blue look distinctly uncomfortable in reverse. Is this just a question of familiarity, or is there a ‘handedness’ in colour distribution as well as in composition?

Martin Kemp

Department of the History of Art, University of Oxford, 59 George Street, Oxford OX1 2BE, UK

The electrifying Australian

How an egg-laying mammal challenged the world's finest zoologists.

Platypus: The Extraordinary Story of How a Curious Creature Baffled the World

by Ann Moyal

Allen & Unwin/Smithsonian Institution Press: 2001. 248 pp. US\$21.95, A\$29.95

David Penny

If any animal has mystery and charisma, it is the platypus. This book tells its story, and it's a winner. In *Platypus*, Ann Moyal uses this extraordinary animal to record the continuing 200-year puzzle of understanding the Australian fauna. Her book has something for everyone — the excitement of a detective story, the history of biological ideas, the frustration at the fact that morphologists appear not to want to understand the evolution of structure in relation to function.

The arrival in England in 1799 of a platypus skin, preserved in rum, highlighted the distinctiveness of the new (to Europeans) Australian fauna. For the next 100 years, every leading zoologist debated the nature of this apparent mixture of mammal, reptile

and bird — Jean-Baptiste Lamarck, Georges Cuvier, Etienne Geoffroy St-Hilaire, Richard Owen and Thomas Huxley all had their opinions, occasionally uncharacteristically humble. It was 25 years before Johann Meckel discovered the presence of mammary glands. "If these glands produce milk, let's see the butter," retorted one disbeliever.

But at least the platypus was now recognized as a mammal. And therefore, went the reasoning, it must give birth to live young. There was no 'sense' in laying eggs if the young were to suckle — even if the aborigines said that the platypus laid eggs. Thousands of platypuses were shot to settle the question. This is the dark side of zoological history, the 'search and destroy' attitude that turned zoological research into the killing fields. One Scottish naturalist, William Caldwell, returned home from Australia with the remains of more than 1,300 echidna (the other egg-laying mammal). While I was an undergraduate, it was said, "the difference between a botanist and a zoologist is that botanists actually like animals". To the classical zoologist, the only good animal was a "dead and dissected" one. But finding that the platypus laid eggs didn't settle the still-unresolved mystery of what selective forces led an egg-laying mammal to develop mammary glands and for its young to start suckling.

Finding out the function of the platypus's bill took even longer. Everard Home, from the Royal College of Surgeons in London, had suggested by 1802 that it was an exploratory organ that replaced sight and sound in the underwater world where the platypus foraged. Yet it was nearly 200 years later that a German-Australian research group reported that the bill detected electric

fields; a freshwater shrimp could be detected ten centimetres away. The 'platypus electric' had been found, and European scientists had at last discovered something that the aborigines did not already know.

As Moyal's book shows, 'big science' is not just a phenomenon of the past 60 years — it was already flourishing at the end of the eighteenth century. Some of the earliest discoveries of the unusual nature of the Australian fauna date from Captain Cook's first voyage to the South Seas in 1768–71. The scientific purpose of this expedition was to observe the transit of Venus in Tahiti. The French also mounted several major expeditions to the Southern Hemisphere, one with two vessels and 23 scientists. Even if England and France were at war, in the Antipodes the scientific expeditions could meet and toast the internationalism of science.

My only quibble with Moyal's version of the scientific thought of this period is that she accepts William Whewell's (1830s) simplistic division of the geological theories of the time into 'catastrophist' and 'uniformitarian' — uniformitarian referring to the idea that change occurs at a uniform rate throughout time.

Charles Lyell objected (in his third volume of *Principles of Geology*, 1830–33) to Whewell's description of his theory as uniformitarian. Much earlier, James Hutton had argued against uniform rates of geological change and, even later, Charles Darwin frequently insisted, in respect of biological evolution, that rates could vary — it was the basic mechanisms of change that were the same. Hutton, Lyell and Darwin denied both catastrophism and uniformitarianism. They each combined the concepts of a long timescale and the continuity of testable mechanisms through time (although with variable rates) to give the viewpoint known as 'actualism'. In one of those twists of fate, the loser (Whewell, a strong opponent of both Lyell and Darwin) laid the framework for writing the history. But it really is the time to give up the idea that they were uniformitarians; orthogenesis fits much better into that category.

But such quibbles aside, everybody will enjoy this book, and the story continues to unfold. We still do not know which features of the platypus are truly ancient and which are more recent specializations that allowed an early mammal to become both an aquatic and a burrowing animal. Read it and enjoy. ■

David Penny is at the Institute of Molecular Biosciences, Massey University, Box 11-222, Science Towers D5-01, Palmerston North, New Zealand.



BETTMANN/CORBIS

MICHAEL MASLAN HISTORIC PHOTOGRAPHS/CORBIS



A world without competition

The Unified Neutral Theory of Biodiversity and Biogeography
by Stephen P. Hubbell

Princeton University Press: 2001. 448 pp.
\$29.95, £19.95 (pbk); \$75, £52 (hbk)

Peter A. Abrams

During the past 50 years, physicists have sought a 'unified theory' from which all major categories of forces could be derived. Any physicist sufficiently intrigued by the title of Stephen Hubbell's book to wade through its pages will, however, come away disappointed. This will be partly a result of the constraints on achieving any generality and unification in the field of ecology, and partly because of the limitations of Hubbell's theory. Nevertheless, Hubbell's book contains many interesting ideas and intriguing new findings, and few practising ecologists will fail to get something out of it.

The 'unification' of the title refers to the fact that this theory seeks to predict both the number of species found in a community and the distribution of the relative abundances of those species. Hitherto, these two aspects of biological communities have been explained independently by most (but not all) theoretical biologists. The number of species and their distribution of abundances are clearly dependent on some of the same processes. An extinction reduces the number of species as well as removing one of the species previously in the category of least abundance. Nevertheless, the 'best' theory for species-abundance distributions might not effectively account for species number, or vice versa.

In any event, vast expanses of ecological endeavour are totally independent of these



Biodiversity: Hubbell's predictions are surprisingly accurate for tree species.

questions of species number and statistical distributions of abundance. A fisheries manager will find nothing here to help predict the sustainable harvest of cod, nor will an entomologist derive anything about the characteristics of an effective natural enemy of a devastating pest. The unification proposed by Hubbell has limited scope.

Attempts to construct unified theories have produced laws that, at least initially, appear to be universally applicable. Thus, our hypothetical physicist reader will be surprised to find that Hubbell's theory is based

on an assumption that the vast majority of ecologists believe to be almost universally false. He assumes that every individual in every species in a biological community is identical, and that the total abundance of all species is fixed. In his favoured version of the theory, each individual of each species has a low probability of mutating into a new species. All changes in distribution and abundance occur because of purely random variation in births, deaths, migration and speciation. This is what makes the theory 'neutral'.

Hubbell has long championed the idea that the many tree species in the tropical forests where he works coexist largely because they are competitively equivalent. Random changes in abundance occur slowly enough for speciation events to have replaced the rare, random extinctions. *The Unified Neutral Theory of Biodiversity and Biogeography* generalizes this idea and derives many consequences from the assumption that all individuals are equal. Unfortunately, decades of experiments studying hundreds of species pairs have identified no conclusive cases of competitive equivalence, so one would expect such an assumption to lead to nonsensical consequences.

The surprising thing about Hubbell's theory is that it is pretty accurate in its predictions of many attributes of ecological communities — particularly the distribution of abundances of tree species. On the other hand, this ability is perhaps not as surprising as it initially appears; the parameters determining the shape of this distribution — speciation rate and number of individuals in the set of connected communities — are largely immeasurable. Most of the fits are achieved by adjusting these parameters (or their product) to minimize the difference between theory and observation.

There are two main resolutions to the apparent contradiction of small-scale experiments and Hubbell's global theory: first, non-neutral theories would be capable of fitting the data at least equally well; second, the assumption of equality is close to correct. But Hubbell largely ignores the first possibility. And his arguments for the second are not presented until the final 15 pages of the book, and I at least found them unconvincing. Graham Bell, another biologist who has recently analysed 'neutral' models, also favours the second alternative. However, work by Steinar Engen and collaborators indicates that alternative models are likely to be equally capable of fitting observed species-abundance relationships.

One of the likely consequences of Hubbell's and Bell's arguments is that there will probably be studies to resolve this issue in the near future. Theory must build on an understanding of simple cases and limiting assumptions, and Hubbell's deductions from 'neutral' assumptions have provided a rigorous basis for future work. At a time when books are often bland reviews of papers published many years before, this book has a high proportion of previously unpublished results and stimulating ideas. It is unlikely to be ignored. ■

Peter A. Abrams is in the Department of Zoology, University of Toronto, 25 Harbord Street, Toronto, Ontario M5S 3G5, Canada.

Fertility from a metabolic viewpoint

On Fertile Ground: A Natural History of Human Reproduction
by Peter T. Ellison
Harvard University Press, 2001. 358 pp.
£16.95, \$27.95

Roger V. Short

Peter Ellison is Professor of Anthropology and Dean of the Graduate School of Arts and Sciences at Harvard. He is best known for his exploitation of methods for measuring reproductive hormones in human saliva, thereby making it possible to carry out non-invasive studies of the reproductive endocrinology of people living in remote parts of the world, such as the Ituri rainforest in Central Africa or the mountains of Nepal. This book therefore describes some of his work on the fertility of traditional human societies.

Unfortunately, *On Fertile Ground: A Natural History of Human Reproduction* does not live up to its grandiose title; it would have been better called "Females, fatness and fertility", as this is its central theme. The opening chapters present a superficial (and sometimes incorrect) account of human reproduction, with strained analogies: "the area at the base of the brain called the hypothalamus and an associated bit of glandular tissue called the pituitary gland that hangs off the base of the brain like a tiny holiday ornament." It is excessively and unnecessarily North American in its outlook and citation of references. For example, pages are devoted to the bizarre and discredited views of Margie Profet, who postulated that menstruation was designed to cleanse a woman's reproductive tract of pathogens. Although Ellison cites references in the bibliography to important articles that completely demolish Profet's hypothesis, he never presents these counter-arguments in the text.

Another major failing of the book is the

almost complete lack of figures or diagrams to leaven a turgid text. There are repeated references to the important sex differences in the shape of the human pelvis: "the pubic bones at the front of the pelvis are proportionately larger in females, giving the top of the true pelvis, or pelvic inlet, a more circular and less heart-shaped outline, while the ischial bones and the sacrum at the bottom of the true pelvis, or pelvic outlet, have spread further apart, increasing the diameter of the outlet and making it more circular in form as well." Oh for an illustration! And it would have been interesting to have had some discussion about when and how this sexual dimorphism has developed.

Ellison devotes a major section of the book to breastfeeding. Unfortunately, there is no mention of the entero-mammary circulation, which supplies specific immunoglobulins in breast milk to counter-act enteric infections in the baby. As diarrhoea is the single greatest killer of babies worldwide, its prevention is a vital piece of our natural history. There is no mention of the inhibitory peptide that accumulates in the alveoli of the mammary gland and turns off milk production if the milk is not removed, thereby making the mammary gland a unique supply-meets-demand organ. There is no mention of the fact that human nipple sensitivity increases greatly on the day of birth, or that in animals, each suckling bout sends afferent neural inputs to the hypothalamus, resulting in the central release of β -endorphin, the 'feel-good hormone', that depresses dopamine secretion and hence elevates prolactin.

Likewise, there is no mention of the fact that in animals, denervation of the teat immediately abolishes the inhibitory effects of lactation on ovarian activity. Instead, Ellison chooses to dismiss the results of P. W. Howie and A. S. McNeilly, who showed in the 1980s that suckling frequency and duration are unquestionably related to the duration of lactational amenorrhoea. In its place, he proposes a "strong alternative hypothesis,"

Fertility symbol: fatness and fecundity have long been linked in the human imagination.

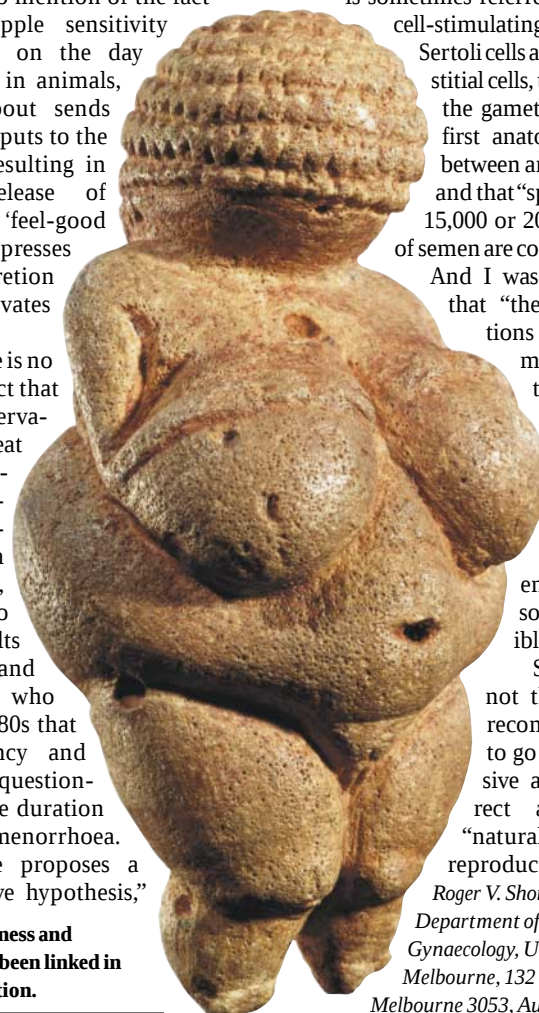
in which "relative metabolic load" is said to be responsible for the lactational inhibition of ovulation. Although we are given no clues as to how the body might sense this "relative metabolic load," or how this information is then used to inhibit ovulation, we are assured that his hypothesis fits the available data better than the nursing frequency hypothesis, and is sounder theoretically! How come that high-yielding dairy cows in negative energy balance, which are milked only twice a day, show little inhibition of ovulation postpartum, whereas low-yielding beef cattle, frequently suckling their calves, have a far longer period of lactational anovulation?

There is a good rebuttal of Rose Frisch's hypothesis that the percentage of body fat is the trigger for the onset of female puberty. But Ellison then goes on to propose an alternative "pelvic size hypothesis," suggesting (without any evidence) that there are mechanical determinants of the timing of puberty. Curiously, there is no mention of the timing of the onset of spermatogenesis in boys, or how it is that boys grow to be considerably taller than girls.

Indeed, the male gets short shrift throughout this book. We are told, quite incorrectly, that follicle-stimulating-hormone in the male is sometimes referred to as interstitial cell-stimulating hormone, that Sertoli cells are also called interstitial cells, that the location of the gametogenic cells is the first anatomical distinction between an ovary and a testis and that "sperm counts above 15,000 or 20,000 per millilitre of semen are considered 'normal'." And I was amazed to learn that "the thermal oscillations produced in the mucin strands by the woman's body heat also resonate with the beating of the sperm cell's tail to aid its progress," although no reference is given to the source of this incredible information.

Sadly, this book is not the place I would recommend anybody to go for a comprehensive and factually correct account of the "natural history of human reproduction". ■

Roger V. Short is in the Department of Obstetrics and Gynaecology, University of Melbourne, 132 Grattan Street, Melbourne 3053, Australia.



Science in culture

Through a looking glass

Did major artists use optical devices to plot points to help them paint?

Rex Dalton

Last year, as artist David Hockney was working on his soon-to-be-published book on lens use in creating paintings, a common fascination with optics brought him together with Arizona physicist Charles Falco. The resulting collaboration may leave a lasting imprint on the worlds of art and science.

Hockney holds, controversially, that before the advent of photography in the mid-nineteenth century, major European artists used optical devices for over 500 years to help capture exactness of expression or perspective in their pictures.

Hockney's best known works include huge collages constructed of hundreds of small photographs, giving him a special eye for creating pictures at some distance with a lens. In 1999, he started to develop his theory of lenses as artists' aids after attending a London exhibition of the works of the early nineteenth-century French painter Jean-Auguste-Dominique Ingres, where he was struck by their extraordinarily confident detail. Like a detective with a refined draughtsman's eye, Hockney then began examining paintings to build a wealth of evidence from the Renaissance to the nineteenth century. Artists he contends used a lens include Caravaggio, Velazquez, van Eyck, Holbein, Leonardo and Lotto.

Hockney's views are controversial on several counts. Ideologically, they offend a traditional notion that the creative process is independent of any form of 'copying'. And on the practical front, there is no documentation of optical devices being used. But Hockney argues that few artists — if any — give away their individual tricks.

With optical science and mathematical formulas, Falco, a condensed-matter physicist at the University of Arizona, Tucson, has helped to affirm many of Hockney's theories. After an article in the *New Yorker* (31 January, 2000) described Hockney's theories, Falco contacted the artist, offering his assistance. In a lengthy series of correspondence — much of which is included in Hockney's book — Falco buttressed Hockney's artistic finds with scientific facts. In particular, Falco showed Hockney that a curved mirror is itself a lens, something not widely understood by art historians. This allowed Hockney and Falco to use mirrors to recreate centuries-old conditions to mimic how paintings may have been created.

For instance, if the geometry of a painted scene and the size of the canvas are known, the focal length of the lens can be calculated. And if the



All done with mirrors: Hockney (left) believes many masters used optical aids; how else could van Eyck achieve such detail in this fifteenth-century portrait?

focal length and the depth of field are known, the lens diameter can be determined. Then a lens can be created to test its use for a painting.

In one of the more vivid examples, Falco helped to dissect an AD 1543 painting by Lorenzo Lotto entitled *Husband and Wife*. In a fax to Hockney, reproduced in the book, Falco calls it "a smoking gun" providing "extremely strong scientific evidence" for Hockney's theory. With assumptions based on measurements of the subjects in the painting, Falco calculated fairly precise properties of a lens for the painting. The focal length of the lens was determined, then the lens's curvature. Such lenses were available more than 100 years before Lotto's painting. Knowing the theoretical properties of the lens, Falco reconstructed it, and used it to recreate distortions in a pattern in a tapestry in the painting, showing how Lotto may have sketched the piece.

Falco, who is an art collector and aficionado of industrial design, particularly classic motorcycles, says he is enjoying the opportunity of a lifetime. Working with Hockney has opened up new research vistas for him in applied optical sciences. The feeling is mutual. "He really gave me a boost," says Hockney of Falco. "He allowed us to confirm things." For a BBC documentary to be shown in October, Hockney went to Florence, Italy, to examine the work of Filippo Brunelleschi,

universally credited with first creating perspective in about AD 1420. "The birth of perspective came from an optical projection," says Hockney, shored up by a re-enactment in Florence. "Art historians have never said that. But the evidence is very strong; the science can't be denied."

Some prominent art historians are impressed by Hockney's theories and writing, although not all are convinced by every example. "It is absolutely fascinating what Hockney is doing," says Martin Kemp of Oxford University. "He is making us look at things afresh." Falco's contribution, he adds, "is most valuable. It is not proof in the strictest sense; but it is supportive and shows no contradictions." But there are doubts among art historians such as Walter Liedtke, a curator of European paintings at the Metropolitan Museum of Art in New York City. "I think Hockney goes too far with his ideas," says Liedtke.

It is important to remember that "each artist has a personal style," notes Liedtke. "We have a brain; the camera or lens does not. Experience and imagination will always intervene." Hockney, of course, agrees on the importance of style. But he feels he is unmasking a secret artistic technique. ■ Rex Dalton is *Nature's West Coast correspondent based in San Diego.*

Secret Knowledge: Rediscovering the Lost Techniques of the Old Masters, by David Hockney will be published in October (US: Viking Studios/Penguin Putnam; UK: Thames & Hudson).

Victor and victim

The true message of *Frankenstein* is about morality, not mad science.

Howard P. Segal

The increasing use of the word 'Frankenfood' by critics of genetically altered food is merely the latest instalment in the unending saga of anti-scientific sentiments allegedly originating with Mary Shelley's *Frankenstein* (1818). It bespeaks the novel's continuing influence that so many immediately recognize the derivation of the term. Yet 'Frankenfood' also echoes persistent distortions of the novel. Those who use *Frankenstein* to bash not just biotechnology but science overall have apparently never read the book, or have never read it carefully.

In truth, *Frankenstein* is hardly a Luddite tract — its message is not a call to destroy laboratories or experiments in the manner of the legendary English machine-breakers of the same period. Nor is its message akin to the famous 10-year moratorium on all scientific research proposed by the Bishop of Ripon in 1927.

Instead, *Frankenstein* insists that scientists must take moral considerations into account before, during and after research and development, and that they must assume responsibility for the outcomes — both intentional and unintentional — of their experiments. Only if experiments prove harmful to society should they stop or be stopped. It is hardly surprising that *Frankenstein* is not more specific here. These now-commonplace positions were barely discussed in 1818, when science was largely the province of wealthy gentlemen whose research was immune from such regulations as university or governmental reviews of experiments on human subjects.

Contrary to the presentation of the story in most films and plays, what actually troubles Shelley about the scientist Victor Frankenstein is not so much his quest to

discover the "cause of generation and life" but rather the secretive, self-centred and finally self-destructive manner in which he pursues this primitive form of physiological engineering. Victor works alone, confides in no one and steadily abandons his family and friends in the attempt to win fame (but not fortune, as he is already well-to-do).

What further disturbs Shelley is not so much Victor's use of dead animals and people to construct his being as his indifference towards how the creature would look, would relate to its creator, and would function outside the laboratory. Victor builds an eight-foot-tall being simply because larger body parts are easier to work with. He never ponders his creature's appearance until he brings it to life.

Indeed, it is critical to the novel's message that the creature is nameless. Contrary again to so many movie and stage versions, the name 'Frankenstein' belongs not to the creature but only to Victor and his family. Abandoned by Victor at 'birth', the unnamed being later compares its miserable plight to God's loving creation of Adam.

Yet it is too easy to characterize Victor as the quintessential mad scientist of most popular treatments. Rather, Shelley portrays him as extraordinarily self-centred. Victor himself notes that he is "not recording the vision of a madman". Instead, from the age of 15 until his death, he repeatedly blames "fate" for all his misfortunes. As Shelley understood, a truly mad scientist might escape moral responsibility for his actions.

Popular culture has also misrepresented the creature as an uncaring monster who kills innocent people without remorse. By now, 'Frankenstein' has nearly become a generic term for monster. In the novel, however, the creature feels guilty about its deeds and ironically, apart from appearance, is far more appealing than Victor. In fact, it is self-educated, sensitive and articulate, unlike the ignorant, grunting beasts of most popular versions. The creature is thereby capable of becoming Victor's missing moral compass. If Victor blames "fate", the creature blames itself, planning to build a bonfire and commit suicide as penance for its actions. Had Shelley believed that scientists should not explore the "cause of generation and life", she would surely have portrayed a genuine moral monster.



Celluloid science: Kenneth Branagh plays Frankenstein alongside Robert de Niro (left) as his monster.



Still, recent feminist perspectives illuminate Shelley's implicit condemnation of Victor's intended usurpation of women's unique reproductive role.

Victor's arrogance towards women is no less outrageous than his abandonment of his creature. His replacement of female procreation constitutes a threat to women far greater than any presented by his creature. In this context, Victor's eventual refusal to satisfy his creature by completing a female companion for it — allegedly to prevent the couple from hurting other humans and from having offspring — is not a sincere ethical stance but rather a self-serving attempt to control both women and the creature.

Like many of her contemporaries, Shelley was fascinated by reports of electrical charges being used to produce temporary signs of life in corpses and by speculation over the potential uses of both electricity and chemistry to create new life forms. One can certainly imagine her endorsing experiments in genetic engineering if they were conducted with the external scrutiny and the moral consideration utterly lacking in Victor's case.

Recent debates about cloning animals and eventually people, about stem-cell research creating and destroying life in the laboratory, and, of course, about 'Frankenfood' all clearly pose challenges to even the most fervent advocates of genetic engineering. Yet it is precisely such scientific and ethical dilemmas that engaged Mary Shelley and that make her original, undistorted *Frankenstein* more timely than ever. ■

Howard P. Segal is in the History Department, University of Maine, Orono, Maine 04469-5774, USA.

***Frankenstein* insists that scientists must take moral considerations into account and assume responsibility for the outcomes of their experiments.**

Reasoning for results

Dennis Bray

Let's start on safe ground. We all agree, surely, that theory — the formulation of hypotheses — is important in biology. Techniques are essential, as is the careful collection of quantitative data. But without ideas to give them shape and meaning, those endless successions of base sequences, expression profiles, electrical recordings and confocal images are as featureless as a plate of tofu. All really big discoveries are the result of thought, in biology as in any other discipline. Allostery, genes, DNA structure, chemi-osmosis, immunological memory, ion channels were all once just a twinkle in someone's eye. And the work of most contemporary research laboratories still takes place within a framework of hypothesis, although practitioners may not always recognize this fact. As Charles Darwin once remarked: "How odd it is that anyone

should not see that all observation should be for or against some view if it is to be of any service."

But assuming that biological theory exists, does it therefore follow that theoretical biology is a distinct and legitimate subject? My guess is that there is less agreement on this proposal; there may even be virulent opposition to it. One of the most common arguments against it is that living systems are so prodigal, so unpredictable and, above all, so historical that any attempt at a grand theoretical framework is doomed to failure. The shape of a leaf in a forest is the product of so many chance events, not only during its growth but also in the aimless blundering of its evolution, that it is inconceivable that it could be predicted from first principles. An all-encompassing theory of biology is no more possible than a predictive theory of other large, inchoate systems, such as the weather, the stock market or London's Heathrow Airport.

Moreover, the argument goes, even in cases in which laws and patterns have been observed in living systems — as in zebra stripes, Hodgkin–Huxley equations and protein coiled-coils — these are not truly biological laws. They belong to mathematics, chemistry or physics, and reflect what living systems have in common with the non-living world, not their own unique biological characteristics.

One can argue with these views, but they are rational and worthy of debate. What is harder to understand is the irrational opposition to theory that often surfaces. Many experimentalists seem to regard theoreticians as carpetbaggers. "We spend months in the lab getting data," they say, "and then along comes some character who doesn't know how to hold a pipette and explains our results to us." Apparently, you're not licensed to theorize unless you put the time in and get the data. This is unfortunate because people are good at different things, and some really enjoy reading papers, juggling possibilities and formulating ideas, even if they can't work a pipette. It is true that certain 'biological theorists' in the past

Theoretical biology

Theory needs to be embraced and to become part of the mainstream of biological research. The quality and accuracy of predictions will then inexorably rise.

have indeed been carpetbaggers, rushing into print with ill-conceived and carelessly prepared ideas, and this gave the subject a bad name. But this is a sociological problem, not a scientific one. The cure is to embrace theoretical work and let it become part of the mainstream of biological research. The quality and accuracy of predictions will then inexorably rise.

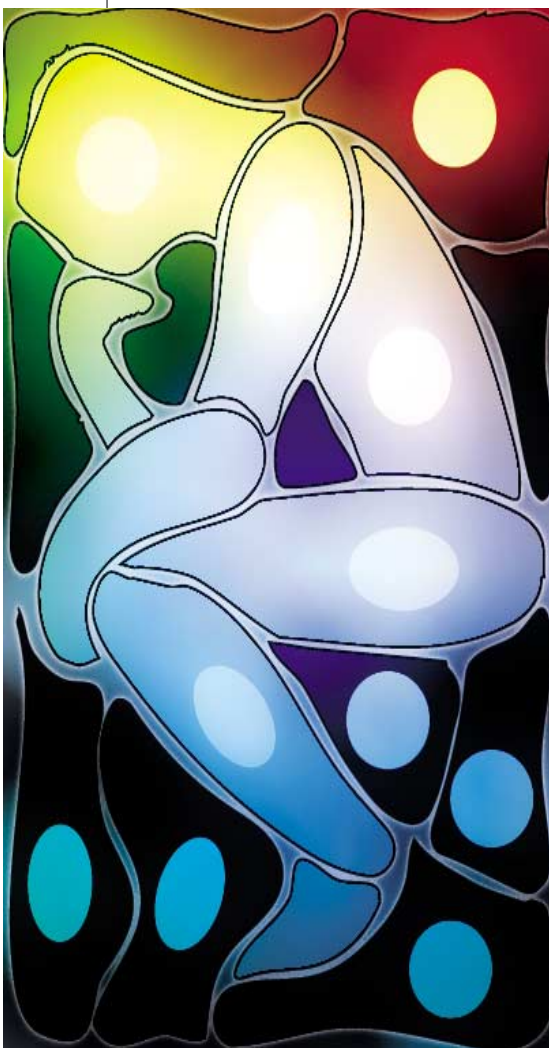
Returning to the chase, the term 'theoretical biology' seems to me just silly — a frozen joke, an oxymoron. Whoever coined the phrase (I'd love to know who it was) was probably being deliberately provocative by drawing parallels with theoretical physics. But although the name is misbegotten, the thing it has come to represent is healthy and growing like a weed. Surfacing briefly in a seminar the other day, I realized that I had no idea whether the traces on the screen were actual electrical recordings made in a physiological experiment or the output of a computer program. It was impossible to tell.

Computer models of action potentials, synaptic integration, heart contraction and even the movements of ions and molecules in cells are now so accurate that they can often be used as experimental objects in lieu of the thing they represent. Biologists can now design and test small genetic circuits in theory and then make them in actual living cells. It seems inescapable that, at least at the level of molecules and cells, biology is moving from an era of data-collection to one of hypothesis-driven research. Progress in this new field will be driven by informed and increasingly quantitative theories — whatever name we choose to give it. ■

Dennis Bray is in the Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK

FURTHER READING

Mayr, E. *This is Biology* (Harvard Univ. Press, Cambridge, MA, 1997).
Noble, D. & Rudy, Y. *Phil. Trans. R. Soc. Lond. A* **359**, 1127–1142 (2001).
Garner, T. S., Cantor, C. H. & Collins, J. J. *Nature* **403**, 339–342 (2000).
Elowitz, M. B. & Leibler, S. *Nature* **403**, 335–338 (2000).
Huszagh, V. A. & Infante, J. P. *Nature* **338**, 109 (1989).
Virtual cell: <http://www.ncrm.uchc.edu>



The biological thinker: making data meaningful.

Some people enjoy reading papers, juggling possibilities and formulating ideas, even if they can't work a pipette.

Achilles' heel of cancer?

Bert Vogelstein and Kenneth W. Kinzler

The p53 protein is inactivated in most human cancers. One outcome is a defect in controlling cell division. Might a virus that exploits this defect prove useful in treating cancer?

A decade or so ago, it was discovered that the p53 protein is subtly mutated, and thereby inactivated, in almost all types of human cancer^{1,2}. These findings raised the exciting possibility that drugs that restore p53's function would be broadly applicable cancer treatments. That hope has not yet been realized, but there have been some imaginative approaches to the problem — the latest of which is described by Raj and colleagues on page 914 of this issue³.

For ten years after its discovery in 1979, the p53 gene was thought to be an oncogene⁴. Under normal circumstances, the protein products of oncogenes stimulate appropriate cell division or interfere with cell death. However, certain mutations in these genes result in their products being switched on all the time, or gaining a new function. This generally leads to inappropriate stimulation of cell growth, resulting in tumours. All commonly used drugs inhibit the ability of their target proteins to bind to other molecules or to catalyse reactions, and so mutated oncoproteins are, in theory, perfect drug targets. Theory has been put into practice with the development of STI-571, a drug that inhibits the aberrantly activated enzyme found in chronic myeloid leukaemia cells⁵.

Yet most genetic alterations in common human cancers affect tumour-suppressor genes, rather than oncogenes. The normal function of a tumour-suppressor gene is to keep cell numbers down, by stopping cells from multiplying or by promoting cell death. When mutated, tumour-suppressor proteins are switched off, and again the result is inappropriate net cell growth. In some cancers, an entire tumour-suppressor gene may be deleted. So tumour suppressors are poor targets for conventional drugs — one cannot inhibit an activity that is not there.

Unfortunately, the same work that identified p53 as a potential target for anticancer drugs also showed that it behaves genetically like a tumour-suppressor gene, not an oncogene¹. One of the main functions of p53 is in the control of cell division. In the presence of damaged DNA or other stresses, p53 prevents cells from progressing from a resting phase of the cell-division cycle (G1 phase) to the DNA-replication phase, or from a second resting phase (G2 phase) to nuclear

Box 1 Using p53 to kill cancer cells

The p53 protein is a tumour suppressor — it keeps cell numbers down by stopping cells from multiplying or by promoting cell death. Loss of p53 occurs in most human cancers, so it would be useful to be able to restore its function. Several innovative strategies have been suggested:

- Introduce normal p53 genes into a cancer cell with mutant p53.
- Introduce a small compound

that converts mutant p53 proteins from an abnormal to a normal shape.

- Add a protein that attaches itself to mutant p53 and kills cells.
- Stimulate the host's immune response to mutant p53 peptides.
- Introduce drugs that disrupt the interaction between the MDM2 or E6 proteins and p53. (MDM2 and E6 negatively regulate p53; they are present at abnormally high levels in

some cancer cells, so 'quench' any normal p53.)

- The strategy described by Raj *et al.*³: introduce the adeno-associated virus, which mimics damaged DNA. Cells with mutant p53 cannot activate the usual p53-dependent 'checkpoint' that is induced by DNA damage, and eventually die.
- Infect cells with viruses that can replicate only in cells without normal p53; the viruses kill these cells. **B.V. & K.W.K.**

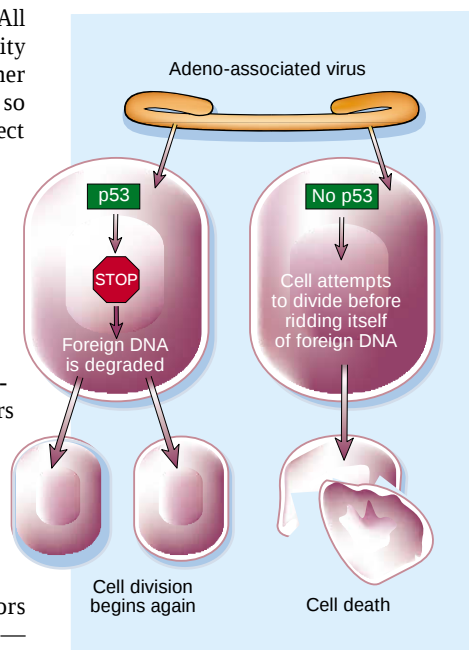


Figure 1 The strategy used by Raj *et al.*³ to kill cells that lack p53. This is one of many approaches being developed to restore p53 function to cancer cells (Box 1).

division (mitosis)^{6,7}. In this sense, p53 is a true "guardian of the genome"⁸, and its loss can have severe consequences under certain conditions.

There are several approaches to restoring normal p53 function (Box 1). The most direct involves gene therapy⁹ to transfer a

wild-type p53 gene to cancer cells that express mutant p53. Although this has shown promise in animals and humans, it is currently impossible to deliver any gene specifically to cancer cells, and externally derived p53 is likely to be as harmful to normal cells as to cancer cells. So gene therapy with p53 will probably be highly toxic unless more specific modes of delivery can be developed.

Other ambitious strategies aim to convert mutant p53 protein into its wild-type form. Normally, one end of the p53 protein (the carboxy terminus) folds back to inhibit its central, DNA-binding region. Peptides have been developed that interfere with this process and in so doing can tickle certain mutant p53 proteins to adopt an active conformation¹⁰. Small compounds that interact directly with the DNA-binding domain and partially activate mutant forms have also been identified¹¹, and sophisticated RNA-based enzymes have been developed for similar purposes¹². Although more potent substances must be developed, this approach is exciting because the compounds should be specific to cancer cells — no normal cell contains mutant p53.

Another angle of attack is provided by the fact that mutant p53 proteins accumulate in cancer cells, whereas normal cells have little p53 unless they are stressed (for example by DNA damage). So, mutant p53 can be decorated with proteins that are toxic only when bound to p53, providing a treatment that would theoretically be harmless to normal cells¹³. It might also be possible to stimulate

patients to produce antibodies that react with cells containing mutant p53 (ref. 14).

Inside cells, p53 is bound to other proteins that negatively regulate it⁴. For example, the MDM2 protein binds to p53 inside the nucleus, enabling p53 to be exported from the nucleus into the hands of a protein-degrading machinery known as the proteasome. In some types of cancer (liposarcomas, for example), p53 is normal but there are too many copies of the MDM2 gene, so there is more MDM2 protein around to inactivate p53. The E6 protein from human papilloma viruses has a similar function, leading to cervical cancer. Drugs that interfere with the binding of MDM2 or E6 to p53 might free p53 from these protein shackles, suppressing sarcomas or cervical cancers¹⁵.

The strategy outlined by Raj *et al.*³ is likewise directed not towards p53 itself, but (in this case) towards the consequences of p53 action (Fig. 1). The authors infected cultured cells with adeno-associated virus (AAV), and found that this activates the p53 checkpoint, presumably because the unusual genome of AAV, which contains single-stranded DNA and DNA hairpins, mimics damaged DNA. Indeed, a DNA molecule corresponding to the AAV hairpin, but which contains no genes, had the same effect³. Cells containing normal p53 paused in the G2 phase to rid themselves of the AAV genome, and then began dividing again. But cells without functional p53 could not maintain their arrest in the G2 phase, and instead began a catastrophic nuclear division that led to cell death. The AAV strategy also proved successful in shrinking experimentally induced cancers in mice.

This approach might be more specific than standard DNA-damaging drugs, which interact with a plethora of cellular constituents and have diverse toxic effects. It would not matter if AAV entered normal cells, because these cells have functional p53. One problem, however, is that the virus is inactivated and non-infectious; because it cannot spread from cell to cell, it must be delivered specifically to almost every cancer cell.

There is a viral-based strategy that potentially solves part of this delivery problem¹⁶. The ONYX-015 adenovirus was originally developed as an anticancer agent that can replicate only in p53-deficient cells, killing them directly. Although the replicative abilities of this virus are more complex than once thought¹⁷, the idea of a live virus that can infect p53-deficient cancer cells but cannot multiply in normal cells is attractive. Infection of only a few cancer cells could theoretically spread to the whole cancer. But there are drawbacks¹⁷. It is not trivial to infect even a small fraction of tumour cells by systemic administration (for example, by injecting adenoviruses intravenously), and the host's immune system may fight back. Immune

reaction to the inactivated AAV might not be such a problem³.

No cancer patient should expect any of these innovative approaches to cure them in the near future. But continued research into the structure of p53, the proteins to which it binds and the pathways that it controls should eventually lead to the maturation of one or more of these strategies — each of which has the potential to benefit millions.

Bert Vogelstein and Kenneth W. Kinzler are at the Howard Hughes Medical Institute and Johns Hopkins Oncology Center, Baltimore, Maryland 21231, USA.

e-mails: vogelbe@welch.jhu.edu

kinzlike@welch.jhu.edu

1. Baker, S. J. *et al. Science* **244**, 217–221 (1989).

2. Nigro, J. M. *et al. Nature* **342**, 705–708 (1989).
3. Raj, K., Ogsten, P. & Beard, C. *Nature* **412**, 914–917 (2001).
4. Vogelstein, B., Lane, D. & Levine, A. J. *Nature* **408**, 307–310 (2000).
5. Mauro, M. J. & Druker, B. J. *Oncologist* **6**, 233–238 (2001).
6. Waldman, T., Lengauer, C., Kinzler, K. W. & Vogelstein, B. *Nature* **381**, 713–716 (1996).
7. Diggle, C. P. *et al. Int. J. Cancer* **93**, 199–203 (2001).
8. Lane, D. P. *Nature* **358**, 15–16 (1992).
9. Roth, J. A. *et al. Semin. Radiat. Oncol.* **10**, 333–342 (2000).
10. Selivanova, G. *et al. Semin. Cancer Biol.* **8**, 369–378 (1998).
11. Foster, B. A., Coffey, H. A., Morin, M. J. & Rastinejad, F. *Science* **286**, 2507–2510 (1999).
12. Watanabe, T. & Sullenger, B. A. *Proc. Natl Acad. Sci. USA* **97**, 8490–8494 (2000).
13. Da Costa, L. T. *et al. Proc. Natl Acad. Sci. USA* **93**, 4192–4196 (1996).
14. Noguchi, Y., Richards, E. C., Chen, Y. T. & Old, L. J. *Proc. Natl Acad. Sci. USA* **92**, 2219–2223 (1995).
15. Hietanen, S., Lain, S., Krausz, E., Blattner, C. & Lane, D. P. *Proc. Natl Acad. Sci. USA* **97**, 8501–8506 (2000).
16. Bischoff, J. R. *et al. Science* **274**, 373–376 (1996).
17. Dix, B. R., Edwards, S. J. & Braithwaite, A. W. *J. Virol.* **75**, 5443–5447 (2001).

Quantum optics

Photons yield to peer pressure

Paul Kwiat

In 1960 the invention of the laser allowed classical light to be amplified inside a cavity. New experiments show that photons in a special quantum state can also be amplified.

Quantum entanglement between two particles means that measuring the behaviour of one instantly determines the behaviour of the other, even when they are physically far apart. Erwin Schrödinger once described this peculiar connection as “the characteristic trait of quantum mechanics, the one that enforces its entire departure from classical lines of thought”. Now, on page 887 of this issue, Lamas-Linares, Howell and Bouwmeester¹ report some of the first basic experiments on amplifying entangled photons through a process analogous to that occurring in lasers. Because entanglement is a key ingredient in many quantum information procedures, such as quantum computing, teleportation and cryptography, a method for reliably producing entangled states of several photons could have significant ramifications, especially for quantum communications using optical techniques.

Entanglement describes a system with several components in which the individual parts carry no information but nevertheless share quantum correlations with each other that are stronger than those allowed by classical physics. For example, photons can be polarized — the polarization describes the oscillation direction of the electric field associated with a light wave. Polarization filters, such as Polaroid sunglasses, will let through photons polarized in one plane but block those polarized at right angles, and so can be used to measure photon polarization. If two photons have entangled polarizations, each photon individually would appear completely unpolarized (with no particular oscillation direction) and yet measuring the

polarization of one completely determines the polarization of the other. It is as if you flipped two coins, each of which was equally likely to come up heads or tails, and yet they always gave the same results — that is, both heads or both tails.

Although normal coins do not behave like this, it has been known for some time how to produce pairs of photons that do display such bizarre quantum-mechanical correlations^{2,3}. Ultraviolet photons sent into a nonlinear optical crystal will sometimes split into two infrared daughter photons (a process known as ‘spontaneous downconversion’), each travelling in its own beam, which are polarization entangled. Unfortunately, the likelihood of each photon splitting in this way is less than 1 in 10 billion. So even though typical experiments use pulses that have several billion ultraviolet photons in them, often not even one will lead to a downconversion pair. The chance of two of them splitting, which can lead to an entangled quartet, is much smaller still: previous experiments have reported four-photon counting rates of only 1 to 2 events per minute^{4–6} (out of 5 billion pulses per minute).

Lamas-Linares *et al.* have significantly increased this rate by using the entangled output of one crystal as a seed for ‘growing’ more entanglement in another crystal (actually the same one, used a second time). Without the seeding the second crystal would produce an entangled beam of about the same brightness as the first crystal. By adding the seed photons, one might naively expect production rates simply equal to the sum of the rates for each process individually (that is, a doubling of the

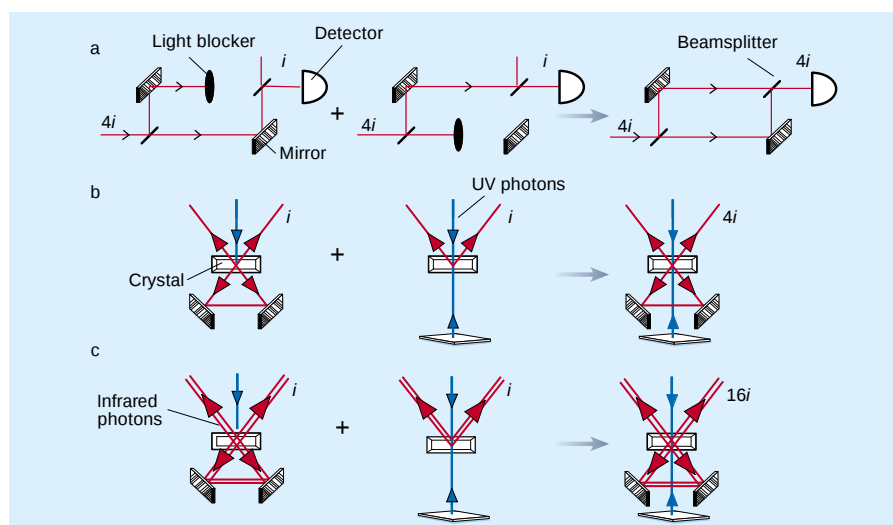


Figure 1 In quantum mechanics, one plus one can equal zero, or four. **a**, An optical interferometer consists of two mirrors and two beamsplitters, which reflect half the light and transmit the rest. If we look at one output of the interferometer, with either path blocked we detect one-quarter of the light sent into the interferometer. If we call the contribution from each path i , we might expect that with both paths open the detector would see $2i$. But in fact the intensity on the detector can range from 0 (if the contributions from the two paths interfere destructively) to $4i$ (if the contributions from the two paths interfere constructively). **b**, In a similar process, infrared downconversion pairs may be produced during the first passage of the ultraviolet (UV) photons through the crystal, or during their return passage. Because these indistinguishable processes can interfere, the final output intensity of pairs can vary from 0 to four times the pair intensity from a single passage through the crystal. **c**, Lamas-Lamare *et al.*¹ show that further enhancement is possible with the four-photon process — going through the crystal twice increases the rate of four-entangled photon production 16-fold.

output of each crystal alone). Not so, thanks to the magic of quantum interference (Fig. 1a) and stimulated emission.

Because the two downconversion processes can interfere, the production of photon pairs can be enhanced by a factor of four (Fig. 1b). The enhancement of downconversion was first demonstrated in 1994 (ref. 7), but the new work extends this to pairs that are polarization entangled. Even more striking is the enhancement effect on four-photon processes (Fig. 1c). The simple interference analogy would again lead one to predict a maximum enhancement factor of four. But the enhancement for four-photon entanglement is higher still — a factor of sixteen — owing to the increased efficacy of stimulated emission as the number of photons increases. Loosely speaking, the photons experience more peer pressure when there are more photons to pressure them into conforming. Lamas-Linares *et al.*'s results clearly show the predicted enhancement, evidence that they have observed the stimulated emission of entangled radiation.

A word of caution: the extra enhancement by a factor of four of the quartet-photon state relative to the twin-photon state should not be misinterpreted to imply that creation of the former is more likely than creation of the latter. Lamas-Linares *et al.* report two-photon counting rates of over 10,000 per second, but four-photon counting rates of only 1 per second. Still, this is a 50-fold increase over the results of initial

experiments done only a few years ago^{4–6}. Equally important, the quality of the four-photon interference has seen a similar improvement — the interference contrast has been raised from about 85% to 97%.

The experiment of Lamas-Linares *et al.* is only a first step toward producing a 'laser' of entangled photons — strictly speaking, it is the analogue of a laser cavity in which the light is cycled just twice, instead of tens to hundreds of times. It will be interesting to see how far the technology can be pushed. Unfortunately, the entanglement is most visible only if a definite number of photons can be identified in each of the two beams. This becomes more and more difficult as the number increases (because of photon loss and detector inefficiency), and could limit the usefulness of the approach. Nevertheless, just as no one knew in 1960 all the possible applications of the optical laser, it can safely be said that the potential of an entangled-photon laser for quantum information applications is also largely unknown.

The authors¹ suggest that this work may lead to some new form of quantum cryptography, with higher transmission rates than are presently available using entangled photon twins^{8–10}. I think this is unlikely, both because any loss will reduce the entanglement, and because the generation rate of these multiple-pair pulses is so much lower than the generation rate of single-pair polarization-entangled photons. Entanglement might also be used to improve the signal-to-



100 YEARS AGO

Mr. W. W. Davis has a paper in "Studies from the Yale Psychological Laboratory"... on some relationships between temperament and effects of exercise. His tests and observations are scarcely sufficient to establish very definite relations, but the conclusions at which he arrives are not without interest. The observations suggest that nervous persons, in training for the development of strength, require light practice, and phlegmatic persons require vigorous practice. The phlegmatic type of temperament is apparently characterised by the presence of much reserve energy of muscle and nerve cell. The nervous type has less reserve energy but a greater ability to use the energy at hand. It is not difficult to apply these principles to practical physical training. They make necessary on the part of the trainer a knowledge, secured either by means of observation or experiment, of the temperament of each man under his charge. From *Nature* 29 August 1901.

50 YEARS AGO

The ground-state of the hydrogen atom is a hyperfine doublet the splitting of which, determined by the method of atomic beams, is 1,420,405 Mc./sec. Transitions occur between the upper ($F = 1$) and lower ($F = 0$) components by magnetic dipole radiation or absorption. The possibility of detecting this transition in the spectrum of galactic radiation, first suggested by H. C. van de Hulst, has remained one of the challenging problems of radio-astronomy. In interstellar regions not too near hot stars, hydrogen atoms are relatively abundant, there being, according to the usual estimate, about one atom per cm. Most of these atoms should be in the ground-state. The detectability of the hyperfine transition hinges on the question whether the temperature which characterizes the distribution of population over the hyperfine doublet — which for want of a better name we shall call the hydrogen 'spin temperature' — is lower than, equal to, or greater than the temperature which characterizes the background radiation field in this part of the galactic radio spectrum. If the spin temperature is lower than the temperature of the radiation field, the hyperfine line ought to appear in absorption; if it is higher, one would expect a 'bright' line; while if the temperatures are the same no line could be detected... We can now report success in observing this line. From *Nature* 1 September 1951.

noise ratio of interferometric devices, such as those used in precision measurements. But at the moment the main value of the research is in giving physicists an ever more complex playground of quantum states; learning how best to use them is the next step. In attempting to build an entangled photon laser (light amplification by stimulated emission of radiation), Lamas-Linares *et al.* have pushed the frontier of exotic quantum-state generation. And what's more, they did it with EASE — entanglement amplification by stimulated emission. ■

Paul Kwiat is in the Department of Physics,

University of Illinois at Urbana-Champaign, 1110 W. Green Street, Urbana, Illinois 61801-3080, USA.
e-mail: kwiat@uiuc.edu

1. Lamas-Linares, A., Howell, J. C. & Bouwmeester, D. *Nature* **412**, 887–890 (2001).
2. Freedman, S. J. & Clauser, J. F. *Phys. Rev. Lett.* **28**, 938–941 (1972).
3. Kwiat, P. G. *et al. Phys. Rev. Lett.* **75**, 4337–4341 (1995).
4. Bouwmeester, D. *et al. Nature* **390**, 575–579 (1997).
5. Ou, Z. Y., Rhee, J.-K. & Wang, L. J. *Phys. Rev. Lett.* **83**, 959–962 (1999).
6. Pan, J.-W. *et al. Nature* **403**, 515–519 (2000).
7. Herzog, T. J., Rarity, J. G., Weinfurter, H. & Zeilinger, A. *Phys. Rev. Lett.* **72**, 629–632 (1994).
8. Jennewein, T. *et al. Phys. Rev. Lett.* **84**, 4729–4732 (2000).
9. Naik, D. S. *et al. Phys. Rev. Lett.* **84**, 4733–4736 (2000).
10. Tittel, W. *et al. Phys. Rev. Lett.* **84**, 4737–4740 (2000).

Cardiovascular biology

Creating unique blood vessels

Peter Carmeliet

When tissues need more oxygen, they release molecules that encourage blood vessels to grow. The discovery of the first such molecule that is specific to one type of tissue has implications for cancer and heart failure.

On a farm, systems of hosepipes are used to deliver vital water to every field. Similarly, in our bodies, blood vessels deliver essential oxygen and nutrients to each organ. New blood vessels form when the endothelial cells that line the inside of the vessels, and the smooth-muscle cells that form a jacket around the outside, grow in response to specific signals. This process is known as angiogenesis and, when derailed, contributes to numerous diseases, including cancer, blindness in diabetics and arthritis¹.

One might assume that it would be irrelevant whether the blood vessels in different organs are alike or not, just as it would not matter what type of hosepipes irrigate the farm, as long as all fields receive the water they require. Yet, beyond their basic need for oxygen and nutrients, different tissues make different demands on their blood vessels. This suggests that, as well as the general signals (such as vascular endothelial growth factor, VEGF) that control angiogenesis in all organs^{2–4}, there should also be tissue-specific angiogenic molecules. On page 877 of this issue, LeCouter and colleagues⁵ describe their discovery of the first such molecule. Their results imply that other tissue-specific angiogenic signals exist. If so, this would create many new — and probably safer — opportunities for stimulating or inhibiting angiogenesis in diseased tissues without affecting healthy organs (Box 1).

Because different tissues have distinct needs, the blood vessels that supply them also differ (Box 1). Take, for instance, the vessels in endocrine glands — organs such as ovaries and testes that produce steroid hormones involved in growth, metabolism, stress, reproduction and sexual development. These hormones must be able to reach

the bloodstream, so vessels in endocrine glands are leaky and their endothelial cells contain fenestrations (tiny gaps through which fluid and small solutes can pass). In contrast, endothelial cells in the brain are

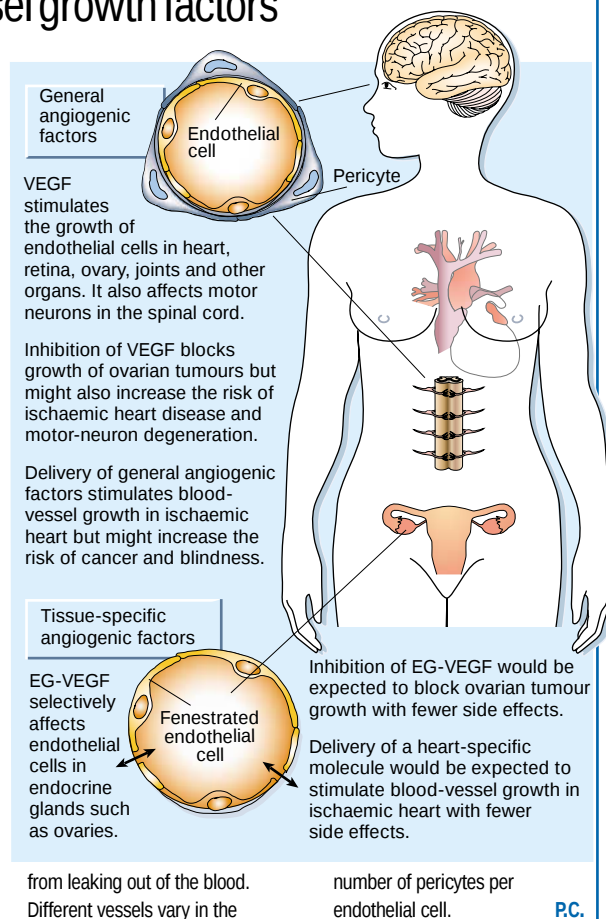
tightly linked to each other and are engulfed by numerous periendothelial cells, which constitute a barrier that protects nerve cells from potentially toxic molecules from the blood.

VEGF is involved in the growth of fenestrated blood vessels. But, given that it is also expressed near non-fenestrated vessels, it cannot be the only factor involved. Moreover, when endothelial cells from endocrine glands are cultured in a dish, they lose their fenestrations. This plasticity suggests that endocrine glands produce signals that are essential to maintain the specificity and growth of their endothelial cells.

LeCouter *et al.*⁵ have now identified such a molecule, which they call endocrine-gland-derived VEGF (EG-VEGF). This molecule functionally resembles and complements VEGF in its ability to induce the formation of endothelial fenestrations. However, EG-VEGF and VEGF are structurally dissimilar and probably work through different receptors. Moreover, VEGF affects endothelial cells non-selectively, and even acts on non-endothelial cells such as motor neurons⁶. EG-VEGF, on the other hand, affects only endothelial cells in endocrine glands. It is likely to be the first member of a large class of tissue-specific vascular growth factors.

Box 1 Blood-vessel growth factors

The properties of general and specific molecules that promote the growth of blood vessels (angiogenesis). Vascular endothelial growth factor (VEGF) is an example of a general molecule, produced by most tissues. Endocrine-gland-derived VEGF (EG-VEGF) is the first tissue-specific angiogenic molecule to be identified⁵; its discovery suggests that there may be others. Tissue-specific angiogenic molecules are needed because blood vessels differ. Bottom, blood vessels in endocrine glands need to be leaky, and have pores known as fenestrations through which hormones, produced by the endocrine glands, can enter the bloodstream. VEGF and EG-VEGF cooperate to induce the growth of fenestrated blood vessels⁵. Top, endothelial cells in the brain are very different: they do not have fenestrations, and are encased in an impenetrable layer of periendothelial cells (pericytes), which prevent potentially harmful molecules



P.C.

These results⁵ provide new insight into how organs develop a blood supply — a question with significance for tissue regeneration and tumour growth. For tumours to grow or tissues to regrow (for example, during the healing of burned skin), they need oxygen, which at first reaches cells by simple diffusion from existing blood vessels. But when the gap between the rapidly multiplying cells and the blood vessels exceeds the limit of diffusion, distant cells become deprived of oxygen (hypoxic). These cells increase their expression of VEGF and other angiogenic molecules by activating hypoxia-inducible transcription factors, so attracting new blood vessels, which restore oxygenation^{1,4,7}. When such hypoxic upregulation of VEGF is impaired, angiogenesis in tumours is reduced⁸ and, surprisingly, motor neurons degenerate⁶.

LeCouter *et al.* show that the expression of EG-VEGF is also upregulated by hypoxia, providing a molecular explanation as to how growing endocrine glands acquire not only more numerous but also more specialized blood vessels. Moreover, because arteries are produced from different precursors to veins, and smooth-muscle cells are generated from distinct precursors in different organs^{4,9}, their results also raise the possibility that tissue-specific angiogenic molecules control these processes, too.

Do tissue-specific blood-vessel regulators contribute to the excessive angiogenesis that occurs in cancers and inflammatory disorders? We don't yet know, but we should perhaps hope that they do. LeCouter *et al.*⁵ show that the forced overexpression of extrinsic EG-VEGF in rat ovaries leads to the production of cysts, with excessive numbers of blood vessels nearby. If the abnormal expression of intrinsic EG-VEGF likewise contributes to angiogenesis in ovarian disorders, then it might be possible to develop drugs that selectively block this process. Inhibition of VEGF does block deregulated ovarian angiogenesis¹⁰ but, because VEGF is non-selective, such drugs might increase the risk of motor-neuron degeneration⁶ and heart failure because of poor blood supply¹. Blocking EG-VEGF would be expected to have fewer side effects.

On the flip side of the treatment coin, LeCouter *et al.*'s results might also have implications for inducing the growth of new blood vessels in blood-deprived (ischaemic) tissues. A good blood supply is crucial; for example, a lack of blood supply to the heart can cause heart failure¹. VEGF and other general angiogenic molecules have been tested for their ability to stimulate angiogenesis in ischaemic heart tissue, and have proved successful in animals.

But the downside of the long-term use of such general molecules is that they might also stimulate angiogenesis elsewhere — in hidden tumours, for example¹¹. Moreover,

although these molecules work in animals, initial results from clinical trials reveal only modest long-term benefit, so it remains unknown whether VEGF alone can induce the necessary formation of stable blood vessels. Additional angiogenic molecules might be needed, such as placental growth factor, which amplifies the activity of VEGF in ischaemic but not normal heart tissue¹². However, this combination has not yet been shown to stimulate angiogenesis. Given that coronary vessels are derived from different precursors to vessels in other tissues, heart-specific angiogenic molecules might be necessary, and safer than more general molecules.

Research into angiogenesis has arrived at the critical stage of translating the findings from studies of mice to humans. LeCouter *et*

al.'s study⁵ has raised the hope that we might be able to achieve this goal more rapidly, and more safely, than we thought. ■

Peter Carmeliet is at the Centre for Transgene Technology and Gene Therapy, Flanders Interuniversity Institute for Biotechnology, University of Leuven, B-3000 Leuven, Belgium.
e-mail: peter.carmeliet@med.kuleuven.ac.be

1. Carmeliet, P. & Jain, R. K. *Nature* **407**, 249–257 (2000).
2. Ferrara, N. *Am. J. Physiol. Cell Physiol.* **280**, C1358–C1366 (2001).
3. Dvorak, H. F., Nagy, J. A., Feng, D., Brown, L. F. & Dvorak, A. M. *Curr. Top. Microbiol. Immunol.* **237**, 97–132 (1999).
4. Carmeliet, P. *Nature Med.* **6**, 389–395 (2000).
5. LeCouter, J. *et al.* *Nature* **412**, 877–884 (2001).
6. Oosthuysen, B. *et al.* *Nature Genet.* **28**, 131–138 (2001).
7. Semenza, G. L. *Curr. Opin. Cell Biol.* **13**, 167–171 (2001).
8. Tszuki, Y. *et al.* *Cancer Res.* **60**, 6248–6252 (2000).
9. Carmeliet, P. *Nature* **408**, 43–45 (2000).
10. Ferrara, N. *et al.* *Nature Med.* **4**, 336–340 (1998).
11. Carmeliet, P. *Nature Med.* **6**, 1102–1103 (2000).
12. Carmeliet, P. *et al.* *Nature Med.* **7**, 575–583 (2001).

Quantum engineering

Protecting the quantum world

Juan Pablo Paz

When quantum systems interact with the environment, classical properties emerge — a process known as decoherence. Although decoherence is unavoidable, it may still be possible to manipulate the outcome.

The world around us looks classical even though the fundamental laws of physics are based on quantum mechanics. At the atomic level, electrons and protons are blurred entities that cannot be described as point-like particles following trajectories. But macroscopic objects have well defined properties: they are either here or there, and not everywhere. So how does the classical world arise from the quantum?

The consensus today is that classical behaviour is an emergent property of quantum systems, induced by their interaction with the environment. This interaction, a fact of life for complex macroscopic objects, is

responsible for the process of decoherence^{1,2}, which is the biggest obstacle to building a viable quantum computer. An attractive idea to control decoherence is to manipulate the environment — so-called environment engineering. As Carvalho *et al.*³ describe in *Physical Review Letters*, a cold trapped ion could be manipulated in this way by applying a number of judiciously chosen laser fields to create an environment in which the stable state of the ion can be chosen (almost) at will.

Decoherence makes most of the states of a quantum system unstable, so that only a small subset of all possible states, the 'pointer

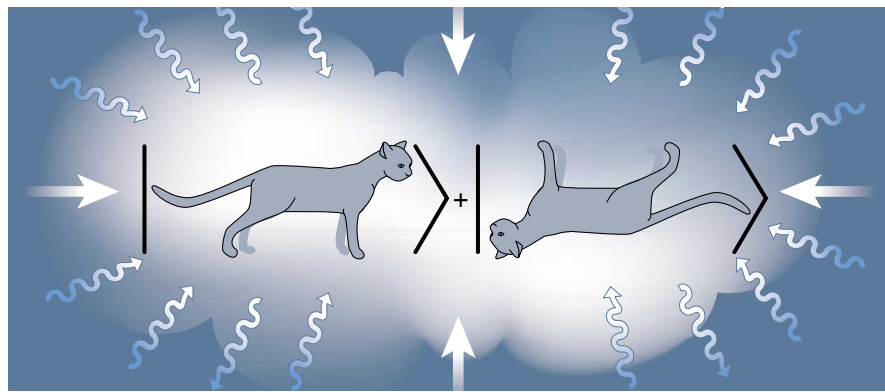


Figure 1 Environment engineering. Quantum states are unstable against the process of decoherence, which is induced by their interaction with the environment. But by using the ideas of environment engineering, physicists can protect some quantum states and prevent their natural degradation. A proposal by Carvalho *et al.*³ suggests that it is possible to do precisely this with the motional state of a cold, trapped ion. This system, when illuminated with an appropriately chosen set of laser fields, can exist in a state that is a small version of the famous dead-and-alive Schrödinger cat state.

states; survive the interaction with the environment. So pointer states, dynamically selected by the environment, are the only ones in which macroscopic systems are allowed to exist. In the real world, pointer states are selected by nature. But in recent years, physicists have found valuable ways to modify the properties of the decoherence process in some systems. They do this by creating an artificial environment that is tailored to control the properties of pointer states.

Ignacio Cirac, Peter Zoller and co-workers⁴ originally devised the idea of environment engineering in 1996. Their work contained the basis for a general method that has been tested by impressive experiments⁵ in which an ion is stored and laser-cooled in a special trap. Ion traps use a particular configuration of oscillating electromagnetic fields to confine the ions to a well-defined region of space, and have been used successfully in high-precision spectroscopy, as well as to demonstrate the feasibility of quantum information processing. The trapped ion can be made to interact with an environment whose properties are designed by the experimenter.

Carvalho *et al.*³ propose a way of preparing pointer states that can persist even in the presence of the form of decoherence that is most effective in disturbing the motion of an ion in a trap: that generated by the interaction with random fields in the trap electrodes. Even with this disturbing interaction, the authors claim their method can be used to select a single pointer state for the centre of mass of a trapped ion, which becomes stable under the effect of the artificial environment. This quantum state can be chosen at will by manipulating the frequency and intensity of the applied laser fields. The method will only protect a single stable state, so rather than specifying several pointer states, which may be relatively stable, the procedure prepares and protects a single specified state in the presence of natural decoherence. The proposal seems to be within the reach of current technology but would require 'stiffer' traps than those used in current experiments.

The method's basic strategy is simple. Suppose that you want to prepare and protect a given quantum state, ψ . The first step of the procedure is to find an operator, D , such that ψ is the only state annihilated by D (that is, $D\psi = 0$). The authors then explain how to design the laser fields, which drive internal transitions of the ions into rapidly decaying states, in such a way that the artificial environment interacts with the ion precisely through the operator D . The key feature of the method is that the net effect of the laser fields is to drive the irreversible decay of the motional state of the ion towards ψ . If the desired state is a given superposition of energy eigenstates (with given coefficients), the method determines the necessary values of the amplitudes of the driving laser fields.

There is a practical limitation to the method. To protect a state in which the ion occupies two positions at once separated by a distance of about 80 nanometres, it is necessary to apply about 15 laser beams to the ion. This is because the number of laser beams is proportional to the number of terms in the expansion of ψ . Increasing the separation into the mesoscopic domain becomes increasingly hard. This kind of quantum state (in which the ion occupies two locations at once) illustrates the most counterintuitive features of quantum mechanics, and is a laboratory cousin of the famous Schrödinger's cat (whose fate is to be in a superposition of alive and dead states). Thus, while the method does not seem to be useful for preparing mesoscopic or macroscopic Schrödinger cat states, it could serve to prepare and protect states of small and tender quantum kittens. The authors also describe in detail how to prepare other interesting quantum states.

A remarkable feature of the debate on the transition between quantum and classical behaviour is that, for the first time in its rather long history, experiments are probing this boundary and playing an important role^{6,7}. Environment engineering, as discussed by Carvalho *et al.*³, is a useful proposal

that will certainly be tested soon in more detail. Moreover, there are other fertile proposals on how to fight decoherence. Many of these ideas, which were born in the context of studies on quantum information processing⁸ (such as quantum error correction), are also being examined experimentally. Most of these are 'controlled decoherence' experiments, which a few years ago appeared to be just thought experiments exploring the nature of the transition from quantum to classical physics. In the near future, a new generation of experiments, in which decoherence is caught in the act, will be able to probe the fuzzy border between the quantum and the classical worlds. ■

Juan Pablo Paz is in the Departamento de Física, Universidad de Buenos Aires, Pabellón 1, Ciudad Universitaria, 1428 Buenos Aires, Argentina.

e-mail: paz@df.uba.ar

1. Zurek, W. H. *Physics Today* **44**, 36–47 (1991).
2. Paz, J. P. & Zurek, W. H. in *Coherent Matter Waves, Les Houches Session LXXII* (eds Kaiser, R., Westbrook, C. & David, F.) 533–614 (Springer, Berlin, 2001).
3. Carvalho, A. R. R., Millman, P., de Mattos Filho R. L. & Davidovich, L. *Phys. Rev. Lett.* **86**, 4988–4991 (2001).
4. Poyatos, J. F., Cirac, J. I. & Zoller, P. *Phys. Rev. Lett.* **77**, 4728–4731 (1997).
5. Myatt, C. J. *et al.* *Nature* **403**, 269–273 (2000).
6. Brune, M. *et al.* *Phys. Rev. Lett.* **77**, 4887–4890 (1997).
7. Haroche, S. *Phys. Today* **51**, 36–42 (1998).
8. Chuang, I. & Nielsen, M. *Quantum Information and Computation* (Cambridge Univ. Press, 2000).

Immunology

Antibody alterations

Alberto Martin and Matthew D. Scharff

In an immune response, antibody molecules are altered so they can bind to intruders more strongly. Some of the molecules that determine which alteration process predominates have now been tracked down.

One of the ways in which the immune system fights off intruders is to produce antibodies, which bind to and neutralize foreign molecules (antigens). The immune system — specifically B cells — must be able to generate enough different antibodies to recognize every possible antigen, so extraordinary antibody diversity is generated before exposure to foreign antigens. But these primary antibodies almost always have low affinity for their targets, and cannot neutralize pathogens or toxins. So, after exposure to an antigen, the variable (V) regions of the antibody genes, which encode the antigen-binding site, acquire many changes — some of which result in higher-affinity binding sites. In some species, this occurs by a process known as gene conversion; in others, by 'somatic hypermutation'^{1–3}. On page 921 of this issue, Sale and colleagues⁴ describe how they used a chicken B-cell line, which normally undergoes gene conversion, to search for factors involved in the diversification of the variable regions. Their results answer some

questions, and raise new ones, about the relationship between gene conversion and somatic hypermutation.

In some species, such as chickens, rabbits, pigs and cows, the diversification of the V region occurs mainly by gene conversion. This process basically involves the acquisition of new DNA sequences, copied from parts of nearby pseudogenes (regions of DNA that are similar to genes but cannot encode a protein) on the same chromosome. In other species, such as sharks, frogs, mice and humans, the V regions of the antibody genes acquire large numbers of single base changes, by somatic hypermutation^{1–3}. It is not clear why these two different processes evolved, and why their species distribution is so sporadic¹. Moreover, the biochemical mechanisms underlying gene conversion and somatic hypermutation have been hard to identify. These questions formed the backdrop for Sale *et al.*'s work⁴.

The authors used the chicken DT40 B-cell line, which is unusual among cultured animal cells because it can undergo high

rates of both homologous recombination and gene conversion^{5,6}. Homologous recombination is the general name given to genetic exchange between two similar DNA sequences; gene conversion is a specific type of homologous recombination. Homologous recombination is also involved in repairing DNA lesions, and studies of yeast have shown that proteins known as RAD proteins are required⁷. In DT40 cells, inactivating the chicken RAD54 protein reduces both homologous recombination and gene conversion, and makes the cells highly sensitive to γ -irradiation⁸, which damages DNA. By contrast, the chicken XRCC2, XRCC3 and RAD51B proteins are also involved in homologous recombination, but have different functions from RAD54 — inactivating them results in only slight increases in cellular sensitivity to γ -irradiation, and greatly increases sensitivity to drugs that crosslink DNA molecules together⁹.

Sale *et al.*⁴ looked at the effects of deleting RAD54, XRCC2, XRCC3 or RAD51B in the chicken B-cell line, and confirmed the need for RAD54 during gene conversion. But loss of the other three proteins had little effect on this process. The implication is that, unlike RAD54, these three proteins are not involved in the intrachromosomal homologous recombination that is required for gene conversion, but participate in other homologous-recombination pathways¹⁰.

The authors suggest that these other pathways might in effect limit the level of somatic hypermutation in normal chicken B cells. They make this proposal because they found that, surprisingly, the deletion of XRCC2, XRCC3 or RAD51B led to very high rates of somatic hypermutation — much higher than the rates of gene conversion. How could this come about? The formation of double- or single-stranded DNA breaks is likely to be required for somatic hypermutation^{11–13}. Such breaks could precede gene conversion, too. Sale *et al.* propose that many more breaks might be produced in the V regions than are revealed by the frequency of gene conversion. Some of these DNA breaks are repaired as a result of gene conversion; others are fixed in normal cells by a different homologous-recombination process (perhaps between matching genes on paired sister chromatids) involving XRCC2, XRCC3 and RAD51B. In cells that lack these proteins, the breaks persist and may trigger the mutational process.

It seems then, as suspected³, that the two mechanisms of V-region diversification can coexist in the progeny of a single B cell. As chicken B cells (and probably B cells from other species) have all of the factors required for both gene conversion and somatic hypermutation, why does one process predominate? This is part of the larger question of how particular sets of DNA-polymerase enzymes and accessory molecules gain

access to the right DNA ends at the right time to carry out such distinct transactions as replication, repair and recombination in all cells¹⁴, and antibody rearrangement and diversification in B cells. As suggested by the effects of inactivating XRCC2, XRCC3 and RAD51B, these events are likely to be regulated by the abundance and activity of the relevant molecules.

For example, as mentioned above, several chicken B-cell lines that carry out gene conversion also display high rates of homologous recombination^{5,6,10}. Perhaps gene conversion predominates over somatic hypermutation in these cells simply because the homologous-recombination enzymes are very active. Alternatively, gene conversion and somatic hypermutation might require different combinations of molecules to be recruited to the V region. Other factors are also likely to contribute, including the accessibility of the pseudogene and of the V region.

Finally, Sale *et al.*'s work⁴ highlights the advantage of working with cultured cell lines: they provide large, relatively homogeneous cellular populations that can easily be genetically manipulated and subjected to detailed biochemical analysis. As the authors point out, the ability to readily inactivate genes in the DT40 cell line by

gene targeting provides a powerful tool. By using it we should be able to discover the full spectrum of genes that are required for gene conversion and somatic hypermutation; to pin down the biochemical mechanisms involved; and to find out exactly how somatic hypermutation is targeted to the V regions of antibody genes. ■

Alberto Martin and Matthew D. Scharff are in the Department of Cell Biology, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, New York 10461, USA.

e-mail: scharff@aecom.yu.edu

1. Diaz, M., Flajnik, M. F. & Klinman, N. *Phil. Trans. R. Soc. Lond. B* **356**, 67–72 (2001).
2. Weill, J. C. & Reynaud, C. A. *Immunol. Today* **17**, 92–97 (1996).
3. Lanning, D., Zhu, X., Zhai, S. K. & Knight, K. L. *Immunol. Rev.* **175**, 214–228 (2000).
4. Sale, J. E., Calandrin, D. M., Takata, M., Takeda, S. & Neuberger, M. S. *Nature* **412**, 921–926 (2001).
5. Buerstedde, J. M. *et al.* *EMBO J.* **9**, 921–927 (1990).
6. Kim, S., Humphries, E. H., Tjoelker, L., Carlson, L. & Thompson, C. B. *Mol. Cell. Biol.* **10**, 3224–3231 (1990).
7. Johnson, R. D. & Jasin, M. *Biochem. Soc. Trans.* **29**, 196–201 (2001).
8. Bezzubova, O., Silbergleit, A., Yamaguchi-Iwai, Y., Takeda, S. & Buerstedde, J. M. *Cell* **89**, 185–193 (1997).
9. Takata, M. *et al.* *Mol. Cell. Biol.* **21**, 2858–2866 (2001).
10. Sonoda, E., Morrison, C., Yamashita, Y. M., Takata, M. & Takeda, S. *Phil. Trans. R. Soc. Lond. B* **356**, 111–117 (2001).
11. Kong, Q. & Maizels, N. *Genetics* **158**, 369–378 (2001).
12. Bross, L. *et al.* *Immunity* **13**, 589–597 (2000).
13. Papavasiliou, F. N. & Schatz, D. G. *Nature* **408**, 216–221 (2000).
14. Sutton, M. D. & Walker, G. C. *Proc. Natl Acad. Sci. USA* **98**, 8342–8349 (2001).

Chemistry

Mystery of an interstellar ion

Annick Suzor-Weiner and Ioan F. Schneider

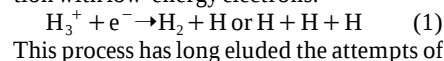
Hydrogen is the most abundant element in the Universe. Its chemistry governs most reactions in space, especially when it is ionized. But the abundance of the H_3^+ ion is not easy to explain.

The elementary molecular ion H_3^+ is a basic building block of interstellar chemistry and star formation. Yet its abundance in diffuse molecular clouds is hard to understand. Much of the debate about H_3^+ has centred on its rate of destruction by interstellar electrons — a rate that varies wildly depending on which calculation or measurement you use. But on page 891 of this issue, Kokouline *et al.*¹ report a new calculation that might reconcile these contradictory results, thanks to the inclusion of a previously overlooked molecular effect.

For such a simple polyatomic molecule, our understanding of H_3^+ has a long and troubled history. It was discovered in the laboratory in 1912 by J. J. Thomson, but it would be many more years before it was seen in space. H_3^+ is usually formed through the exothermic reaction $H_2 + H_2^+ \rightarrow H_3^+ + H$, so it is found wherever molecular hydrogen is present and partly ionized, such as most of the cool regions of the Universe. Being fairly reactive, H_3^+ initiates most of the interstellar chemical reactions in molec-

ular clouds. It is a generous proton donor, and by reacting with neutral molecules ($H_3^+ + X \rightarrow H_2 + HX^+$) it triggers chains of ion–molecule reactions leading to the formation of large organic and prebiotic molecules. So the abundance of H_3^+ is a crucial parameter for models of interstellar molecule formation.

Following the discovery of molecular clouds in the 1940s, astronomers searched in vain for signs of interstellar H_3^+ . It wasn't until the late 1990s that T. Oka and his group successfully detected the infrared spectrum of H_3^+ in dense² and then diffuse³ interstellar clouds of molecules. This long search turned out to be the first of many hurdles in working out the abundance of interstellar H_3^+ . In diffuse interstellar clouds — which contain about 100 particles per cm^3 at a temperature around 100 K and are partly ionized by cosmic rays — the main process that destroys H_3^+ is its dissociative recombination with low-energy electrons:



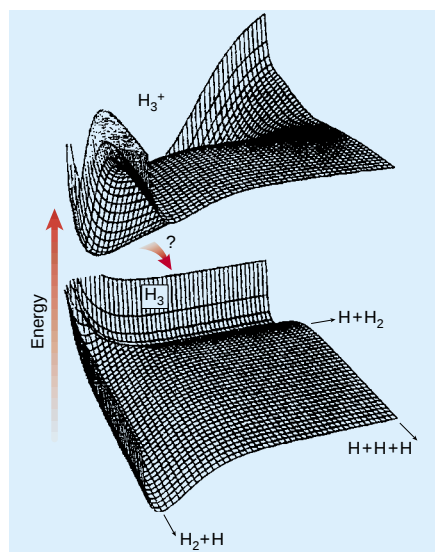


Figure 1 Bridging the gap between theory and experiment. Values for the destruction of interstellar H_3^+ ions by low-energy electrons (dissociative recombination) differ widely between calculations and measurements. The potential energy surfaces for the ground states of H_3^+ and H_3 molecules are separated by a large energy gap, which makes the dissociative recombination of H_3^+ difficult to explain by usual mechanisms. Kokoouline *et al.*¹ include a previously ignored decay mechanism in their new calculation, which allows the efficient fragmentation of H_3 (formed from $\text{H}_3^+ + \text{e}^-$). (Figure reproduced from ref. 9.)

experimentalists, theoreticians and astronomers to pin it down. First, the experimental value for the rate of H_3^+ destruction by low-energy electrons has ranged over several orders of magnitude, depending on the method used to measure it. The most recent measurements converge towards a relatively fast rate constant, but the ion-storage-ring value^{4,5} of about $10^{-7} \text{ cm}^3 \text{ s}^{-1}$ still differs by one order of magnitude from the result⁶ obtained using a flowing-afterglow Langmuir probe ($10^{-8} \text{ cm}^3 \text{ s}^{-1}$).

Second, using the higher destruction rate from the storage-ring data, astrophysical models³ cannot explain the large abundance of H_3^+ ions observed in diffuse interstellar clouds, which have high electron density. As a result of the rapid destruction of H_3^+ , the measured abundance would require stable molecular clouds to be unreasonably large. At most, present astrophysical models can accept the slower rate constant of $10^{-8} \text{ cm}^3 \text{ s}^{-1}$.

Third, reaction (1) has long been considered theoretically impossible for the very cold electrons in the interstellar medium. In the conventional 'Born–Oppenheimer' approximation of molecular physics, which treats separately the fast electronic processes and the slow dynamics of the heavy nuclei, an efficient fragmentation of the $\{\text{H}_3^+ + \text{e}^-\}$ complex is expected to occur only if it has a similar electronic energy to a dissociative

state of the neutral H_3 . But the potential energy well of the H_3^+ ground state lies far above the repulsive potential surface of the neutral H_3 ground state, the only path to dissociation at very low energy (Fig. 1).

The gap between theory and experiment began to narrow when theorists realized (from studies on diatomic ions⁷ such as HeH^+) that dissociative recombination can be helped by dynamical couplings between electronic and nuclear motion, beyond the Born–Oppenheimer approximation. The reaction can then proceed in steps, through a series of H_3 bound excited states, whose potential energy surfaces mimic that of the H_3^+ ground state and span the energy gap between it and the dissociative H_3 . A two-dimensional calculation⁸ including this mechanism led to an H_3^+ destruction rate close to $10^{-9} \text{ cm}^3 \text{ s}^{-1}$ at very low temperature. This value is acceptable for astrophysical models but is still far below the rates observed experimentally.

Kokoouline and colleagues¹ take a decisive step towards resolving this puzzle by identifying a previously neglected decay mechanism. This mechanism also involves a coupling between electronic and nuclear motion, but requires a full three-dimensional treatment of the nuclear dynamics. The Jahn–Teller effect results from the combined vibration and rotation of the three hydrogen nuclei when distorted by the incoming electron from their equilibrium positions in the ground state of H_3^+ . The authors show that this symmetry-breaking effect induces a cascade of energy-reducing transitions that may raise the rate of reaction (1) up to a value of $10^{-8} \text{ cm}^3 \text{ s}^{-1}$, compatible both with recent astrophysical models and Langmuir-probe experiments. Moreover, the effect should increase for rotationally hot H_3^+ target ions, perhaps explaining the larger value measured in storage-ring experiments. The new calculation¹ not only reproduces the tendency for interstellar H_3^+ to break up into three rather than two fragments, but also correctly predicts the vibrational energy of those fragments. A complete quantitative treatment is still required, but the main destruction mechanism for interstellar H_3^+ has now almost certainly been identified. ■

Annick Suzor-Weiner is at the Laboratoire de Photophysique Moléculaire, Université Paris-Sud, Orsay, France.

e-mail: annick.suzor-weiner@ppm.u-psud.fr

Ioan F. Schneider is in the Department of Physics, Université du Havre, France.

1. Kokoouline, V., Greene, C. H. & Esry, B. D. *Nature* **412**, 891–894 (2001).
2. Geballe, T. R. & Oka, T. *Nature* **384**, 334–337 (1996).
3. McCall, B. J., Geballe, T. R. & Oka, T. *Science* **279**, 1910–1913 (1998).
4. Larsson, M. *et al.* *Phys. Rev. Lett.* **70**, 430–433 (1993).
5. Larsson, M. *Annu. Rev. Chem.* **48**, 151 (1997).
6. Glosik, J. *et al.* *Chem. Phys. Lett.* **331**, 209–214 (2000).
7. Guberman, S. L. *Phys. Rev. A* **49**, R4277 (1994).
8. Schneider, I. F., Orel, A. E. & Suzor-Weiner, A. *Phys. Rev. Lett.* **85**, 3785–3788 (2000).
9. Murrell, J. N., Carter, S., Farantos, S. C., Huxley, P. & Varandas, A. J. C. *Molecular Energy Functions* (Wiley, Chichester, 1984).

Daedalus

A cosmic background

One of the most significant discoveries in cosmology is the cosmic microwave background, or 'relic radiation' as the Russians call it. Originally, in the early Universe, it was visible light. But the Universe has expanded so much since then that its frequency has been shifted down all the way to the microwave band, a frequency drop of perhaps 10^5 or so.

Last week Daedalus proposed a directional neutrino telescope, underground as usual to filter out competing particles, but consisting of a tube many metres long. It would be filled with a liquid such as dry-cleaning fluid or heavy water, as used in dedicated neutrino detectors. For some liquids it would be covered with photomultipliers. There would be several telescopes. Giving each some spectroscopic resolution, they could even estimate neutrino energies. Each tube would be sensitive to neutrinos mainly in its long direction, in which they would traverse it from end to end, and so had the best chance of interacting with its contents.

Seeking new uses for his device, Daedalus now reckons that the early Universe should also have created plentiful neutrinos, by the combination of protons or the commonest helium nuclei. As the Universe expanded, these 'relic neutrinos' will have lost energy. By now they should have formed a fairly uniform cosmic background. His directional neutrino telescope will therefore slowly establish that background, and distinguish it from that of the Sun.

It will take many years to sample this very weak neutrino background. But there are very few pieces of information in cosmology (the microwave background is one of them), so the slow effort seems worthwhile. It might even identify a neutrino flux from the nearer stars — Sirius is perhaps the best bet, although Alpha Centauri is nearer. But Daedalus will aim his device in the stellar gaps, especially out of the galactic plane, so as to acquire a true background. There might be surprises, of course. All previous neutrino astronomy has used detectors with no directionality, so it would be intriguing to discover that many neutrinos were coming from some concentrated source.

Indeed, Daedalus would like his telescope to found a whole new branch of astronomy. But he freely admits that the very low rate of data acquisition will greatly restrict the rate at which DREADCO can transform cosmology.

David Jones

Mother's voice recognition by seal pups

Newborns need to learn their mother's call before she can take off on a fishing trip.

In gregarious mammals, mother and pup need to be able to recognize each other's voices in order to be reunited in a crowd¹, a skill that has only been studied in domesticated species where mother and young stay together during the rearing period². In otariids, females have to leave their newborn to feed at sea, but offspring nevertheless develop long-term recognition of their mother's voice³. Here we show that pups of the subantarctic fur seal (*Arctocephalus tropicalis*) can acquire this ability when they are just 2–5 days old, and that the mother times her departure accordingly.

On Amsterdam Island in the Indian Ocean, fur seals are born from late November to early January⁴. The rearing period lasts for an average of 10 months and consists of foraging trips at sea that last for 2–3 weeks at a time, interspersed with suckling periods ashore lasting for 3–4 days⁵. After parturition, mothers usually stay with their young for about a week before leaving the colony on their first feeding trip⁵. When she returns, a female must relocate her own pup among the many similar ones in the rookery⁶. This

reunion depends mainly on vocalizations (Fig. 1) because pups respond specifically to their mother's voice⁷.

We investigated the age at which this specific response develops in pups. By using playback experiments for several days after birth, we assessed the evolution of pups' behavioural responses to their own mother's calls and to those of strange females. As the pup needs to be able to recognize its mother when she comes back from foraging at sea, its response to her voice should have developed before her first departure, which is presumably delayed by the mother until her pup can recognize her.

Responses to playback tests revealed that pups ($n=9$) react vocally to any female's calls a few hours after birth, but after 2–5 days they can respond specifically to their mother's voice. This response was always established before the female's departure, which occurred 2–10 days after parturition. Moreover, the departure date depends on the timing of this voice recognition by the pup ($n=9$; regression analysis, reciprocal- X model: departure date is 10.9–11.9/specific-response date (days after birth); $P<0.05$).

To test the efficiency of the pups' learning, we measured the time taken by pups to find their mother after her return from her first sea trip. Of the monitored mother–pup pairs ($n=15$), 66% met up within 7 min, and the remaining pairs were united within 11 min (regression analysis, double-reciprocal model: percentage of successful meetings calculated as $1/(-0.000356 + 6.45/\text{time required})$; $P<0.01$).

In colonial mammals, the ability of pups to recognize their mother's voice is

crucial for their survival as, in most species, females only feed their own offspring⁸. Mother–young recognition must therefore develop as soon as it is needed, before the first separation. In several bird species, vocal identification between parents and young generally coincides with offspring mobility⁹. In *A. tropicalis*, factors other than call recognition may be involved¹⁰ — for instance, females may leave the shore when pups stop suckling¹¹. Our results show that, in spite of the variable duration of ontogeny, the mother's departure date is linked to the pup's ability to recognize her voice. This supports the idea that recognition of a mother's call by her pup is an important factor in allowing her to go to sea.

Isabelle Charrier*†, Nicolas Mathevon*‡, Pierre Jouventin†

*Laboratoire de Biologie Animale, Université Jean Monnet, 42023 Saint-Etienne cedex 2, France

e-mail: mathevon@univ-st-etienne.fr

†CEFE CNRS UPR 9056,

34213 Montpellier, France

‡NAMC CNRS UMR 8620, 91400 Orsay, France

- Hofer, M. A. in *Parental Care in Mammals* (eds Gubernick, D. J. & Klopfer, P. H.) 77–115 (Plenum, New York, 1981).
- Addae, P. C., Awotwi, E. K., Oppong-Anane, K. & Oddoye, E. O. K. *Appl. Anim. Behav. Sci.* **67**, 77–88 (2000).
- Insley, S. J. *Nature* **406**, 404–405 (2000).
- Kerley, G. I. H. *J. Zool. Lond.* **205**, 315–324 (1985).
- Georges, J. Y. & Guinet, C. *Ecology* **81**, 295–308 (2000).
- Riedman, M. *The Pinnipeds: Seals, Sea Lions and Walruses* (Univ. California Press, Berkeley, 1990).
- Roux, J. P. & Jouventin, P. *NOAA Technical Report NMFS* **51**, 95–102 (1987).
- Stirling, I. J. *Austr. Mam. Soc.* **1**, 389–391 (1975).
- Jones, I. L., Falls, J. B. & Gaston, A. J. *Anim. Behav.* **35**, 1405–1415 (1987).
- Georges, J. Y. & Guinet, C. *J. Zool. Lond.* **251**, 277–287 (2000).
- Gentry, R. L. & Kooyman, G. L. *Fur Seals: Maternal Strategies on Land and at Sea* (Princeton Univ. Press, Princeton, 1986).



Figure 1 A subantarctic fur seal (*Arctocephalus tropicalis*) mother on her return from a foraging sea trip. The mother starts calling soon after arriving ashore and her pup calls in reply to indicate its position within the colony. As they approach one another, they continue to emit vocalizations until they are reunited.

Palaeontology

The beaks of ostrich dinosaurs

Primitive ornithomimids, a ubiquitous group of specialized Cretaceous dinosaurs nested within a clade of predominantly carnivorous theropods, are known to have had teeth, whereas derived ornithomimids had an edentulous beak, which has prompted speculation about their dietary habits¹. Here we describe two new ornithomimid specimens in which soft-tissue structures of the beak have been preserved. These creatures probably used their beaks to strain food sediment in an aqueous environment, rather than for

predation on large animals.

Specimen RTMP (Royal Tyrrell Museum of Palaeontology) 95.110.1 is an almost complete skeleton of the Campanian ornithomimid *Ornithomimus edmontonicus*² from Dinosaur Provincial Park, Alberta, Canada (Fig. 1a). This specimen preserves traces of the keratinous covering (rhamphotheca) on the beak on the tips of the right premaxilla and dentary. As preserved, the soft tissue extends up to 3.5 mm from the rostral and mandibular margins, and fills a gap left by the anterior divergence of the premaxilla and dentary. The rhamphotheca on each jaw tapers posteriorly and terminates just behind the midpoint of the right naris. The upper rhamphotheca extends about 1 mm farther

posteriorly than the lower, and is also slightly longer anteriorly. The surfaces of the dentary and the premaxilla are densely pitted next to the preserved soft tissue, indicating extensive neurovascularization in the region of the ramphotheca.

The other specimen, IGM (Institute of Geology, Mongolia) 100/1133, is a sub-adult of *Gallimimus bullatus*³ from Tsaagan Khushu, Nemegt Basin, Mongolia (Fig. 1b). Although slightly crushed, the skull is remarkably complete. The interior surface of the ramphotheca is preserved on both the right premaxilla and dentary. Small, separate columnar structures extend sub-perpendicularly to the buccal margin of the rostrum. These structures, which are remnants of the lingual surface of the beak, project up to 5.6 mm, as preserved, at the tip of the premaxilla, but are shorter posteriorly. They are tightly packed and evenly spaced (about 0.5 mm from centre to centre). Presumably, these columnar structures would have been at least partially covered laterally by an outer ramphotheca, as in TMP 95.110.1 and extant anseriforms.

These specimens demonstrate the pres-

ence of a beak, presumably keratinous, in ornithomimids (Fig. 2). The *Gallimimus* specimen shows structures (Fig. 1c) that are reminiscent of the lamellae on the beaks of anseriforms⁴. Most anseriforms feed by manipulating food and by straining sediment with their beaks. The lamellae help in segregating food articles from other material, which is expelled from the side of the beak during feeding, and they also aid in cutting plant material during grazing.

In an analysis of 17 diverse modern anseriform species, some straining was observed in all species except the piscivorous merganser⁵, which has large, tooth-like lamellae. The *Gallimimus* condition closely resembles that of the shoveller *Anas clypeata*, an extreme filter-feeding strainer with a diet of plants, freshwater molluscs, ostracods and foraminiferans⁵. Other features of the ornithomimid skeleton have been used as grounds to suggest that they were not carnivorous⁶.

Ornithomimids with numerous small gastroliths have been reported from deposits in China⁷. Although gastroliths are patchily distributed in Dinosauria, they



Figure 2 Artist's rendition of a feeding ornithomimid, showing the sediment-straining method probably used by these dinosaurs.

are most often associated with herbivorous (*Psittacosaurus mongoliensis* and sauropods) or putatively herbivorous (*Caudipteryx zhoui*) groups. Although there is not enough evidence to indicate that these ornithomimids behaved similarly to anseriforms, it is unlikely that these delicate features were used for predation on large animals. Furthermore, ornithomimids are abundant in mesic environments (such as Nemegt⁸, Iren Dabasu⁹ and Dinosaur Park Formations¹⁰), but are rarer in more arid environments (such as Djadokhta^{8,11}). This suggests that ornithomimids may be ecologically tied to food supplies in wetter, more mesic environments, which would be consistent with their beak morphology.

Mark A. Norell*, Peter J. Makovicky*†, Philip J. Currie‡

*American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024-5192, USA

e-mail: norell@amnh.org

†Field Museum, 1400 South Lake Shore Drive, Chicago, Illinois 60605-2496, USA

‡Royal Tyrrell Museum of Palaeontology, Drumheller, Alberta T0J 0Y0, Canada

- Osmólska, H. in *Encyclopedia of Dinosaurs* (eds Currie, P. J. & Padian, K.) 499–503 (Academic, New York, 1997).
- Sternberg, C. A. *Can. Field-Nat.* **47**, 79–83 (1933).
- Osmólska, H., Roniewicz, E. & Barsbold, R. *Palaeontol. Polonica* **27**, 103–143 (1972).
- Zweers, G. A., Gerritsen, A. F. & van Kraneburg-Voogd, P. J. in *Contributions to Vertebrate Morphology* vol. 3 (eds Hecht, M. K. & Szalay, F. S.) 1–109 (Karger, Basel, 1977).
- Goodman, D. C. & Fisher, H. I. *Functional Anatomy of the Feeding Apparatus in Waterfowl (Aves: Anatidae)* 1–193 (South Illinois Univ. Press, Carbondale, 1962).
- Osborn, H. F. *Am. Mus. Bull.* **35**, 733–771 (1916).
- Kobayashi, Y. et al. *Nature* **402**, 480–481 (1999).
- Jerzykiewicz, T. & Russell, D. A. *Cretaceous Res.* **12**, 345–377 (1991).
- Currie, P. J. & Eberth, D. A. *Cretaceous Res.* **14**, 127–144 (1993).
- Eberth, D. A. in *Encyclopedia of Dinosaurs* (eds Currie, P. J. & Padian, K.) 379–384 (Academic, New York, 1997).
- Makovicky, P. J. & Norell, M. A. *Am. Mus. Novitates* **3247**, 1–16 (1998).

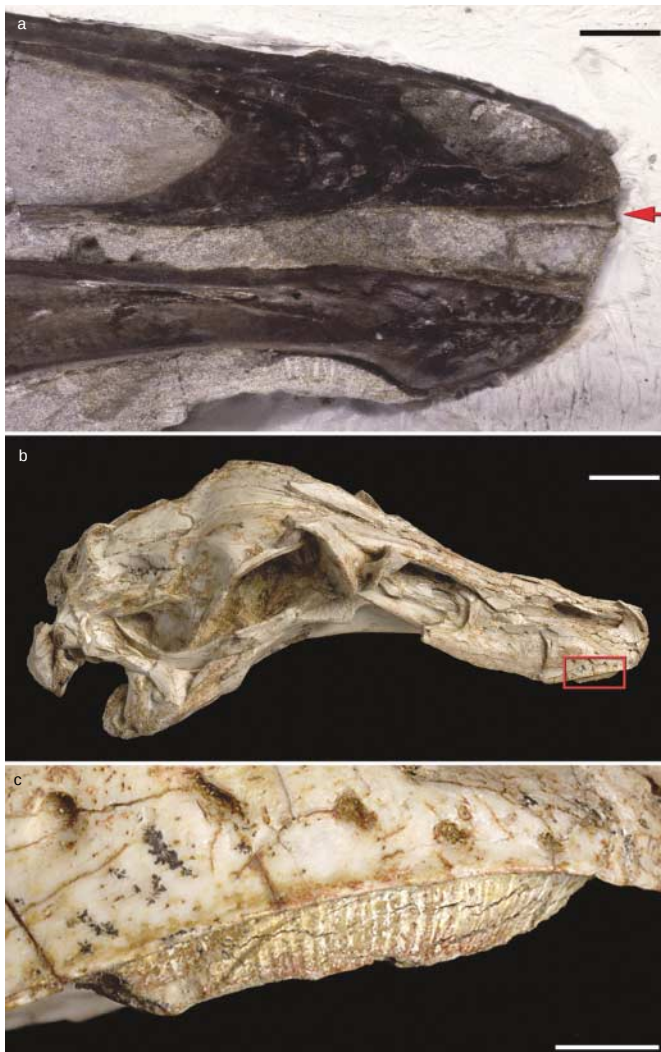


Figure 1 Two ornithomimid specimens, showing preservation of soft-tissue structures of the beak. **a**, The anterior end of the rostrum of *Ornithomimus edmonticus* (RTMP 95.110.1) in lateral view (reversed). Red arrow indicates preserved soft tissue. Scale bar, 1 cm. **b**, *Gallimimus bullatus* (IGM 100/1133) in right lateral view. Scale bar, 2 cm. **c**, Detail of red box in **b**, showing soft-tissue lamellae along the rostral margin. Scale bar, 3 mm.

Nanostructures

Self-assembled domain patterns

The ordered domain patterns that form spontaneously in a wide variety of chemical and physical systems^{1,2} as a result of competing interatomic interactions can be used as templates for fabricating nanostructures. Here we describe a new self-assembling domain pattern on a solid surface that involves two surface structures of lead on copper. The evolution of the system agrees with theoretical predictions^{3–5}, enabling us to probe the interatomic force parameters that are crucial to the process.

The two structures showing this remarkable behaviour are a disordered Pb/Cu surface alloy, produced by deposition of lead vapour onto clean Cu(111), and a lead-overlayer phase, which is produced by continued deposition after the alloy structure has covered the surface^{6–8}. Figure 1a–g shows a sequence of low-energy electron microscope (LEEM) images in which the overlayer structure (bright) grows over the surface alloy structure (dark) during lead deposition at a temperature of 673 K. The most striking feature of this sequence is the evolution of a

pattern from circular islands (average diameter, 67 nm) to stripes and then to circular holes within the lead-overlayer matrix.

This sequence of domain patterns, referred to as droplets, stripes and inverted droplets, respectively, has been predicted to be a general property of domain formation on solid surfaces^{3–5,9,10}. To our knowledge, these results represent the first unambiguous observation of the expected sequence of domain patterns with changing area fraction during deposition, and agree with the predicted positions⁵ of the droplet-to-stripe phase transitions for short-range attractive and long-range dipolar interatomic interactions on solid surfaces.

One reason why these domain patterns form in this system is the high mobility of the islands (Fig. 1a), which contain many thousands of atoms and can move by as much as a micrometre before being incorporated into the more tightly packed droplet structure (Fig. 1b). This mobility allows patterns to form that have hundred-nanometre periodicity and allows the underlying interatomic interactions to be determined.

As the islands move in the force fields of the other islands, the island trajectories contain information about the force fields. Droplet trajectories are consistent with

the expected long-range dipolar repulsion. Because magnetic interactions are not relevant in this system and we can rule out electrostatic interactions from measurements of work-function differences, we propose that elastic interactions arising from a surface-stress difference between the alloy and overlayer structures are responsible for the stabilization of domain patterns^{9,10}.

The behaviour of the patterns at fixed coverage and varying temperature is also qualitatively consistent with theory³. For both droplets and stripes, the feature size decreases with increasing temperature. The most likely reason for this is a reduction in the domain-boundary free energy with increasing temperature. For droplets and inverted droplets, reducing the temperature also allows improvement in the long-range order. Figure 1h is a LEEM image showing the degree of long-range order that can be obtained in the inverted-droplet phase.

If self-assembled structures are to be useful as nanostructure templates, two-dimensional, self-assembled patterns need to be stable at room temperature and resistant to air exposure. This is the case for our Pb/Cu system. Features that correspond to lead-overlayer droplets can be clearly seen in atomic-force-microscopy images taken after cooling to room temperature and 2 hours exposure to air (Fig. 1i).

Richard Plass*, Julie A. Last*, N. C. Bartelt†, G. L. Kellogg*

*Sandia National Laboratories, Albuquerque, New Mexico 87185-1415, USA
e-mail: gkellogg@sandia.gov

†Sandia National Laboratories, Livermore, California 94551-0969, USA

1. McConnell, H. M. *Annu. Rev. Phys. Chem.* **42**, 171–195 (1991).
2. Seul, M. & Andelman, D. *Science* **267**, 476–482 (1995).
3. Hurley, M. M. & Singer, S. J. *Phys. Rev. B* **46**, 5783–5786 (1992).
4. Sagui, C. & Desai, R. C. *Phys. Rev. E* **49**, 2225–2244 (1994).
5. Ng, K.-O. & Vanderbilt, D. *Phys. Rev. B* **52**, 2177–2183 (1995).
6. Henrion, J. & Rhead, G. E. *Surf. Sci.* **29**, 20–36 (1972).
7. Meyer, G., Michailov, M. & Henzler, M. *Surf. Sci.* **202**, 125–141 (1988).
8. Nagl, C., Haller, O., Platzgummer, E., Schmid, M. & Varga, P. *Surf. Sci.* **321**, 237–248 (1994).
9. Marchenko, V. I. *JETP Lett.* **33**, 381–383 (1981).
10. Alerhand, O. L., Vanderbilt, D., Meade, R. D. & Joannopoulos, J. D. *Phys. Rev. Lett.* **61**, 1973–1976 (1988).

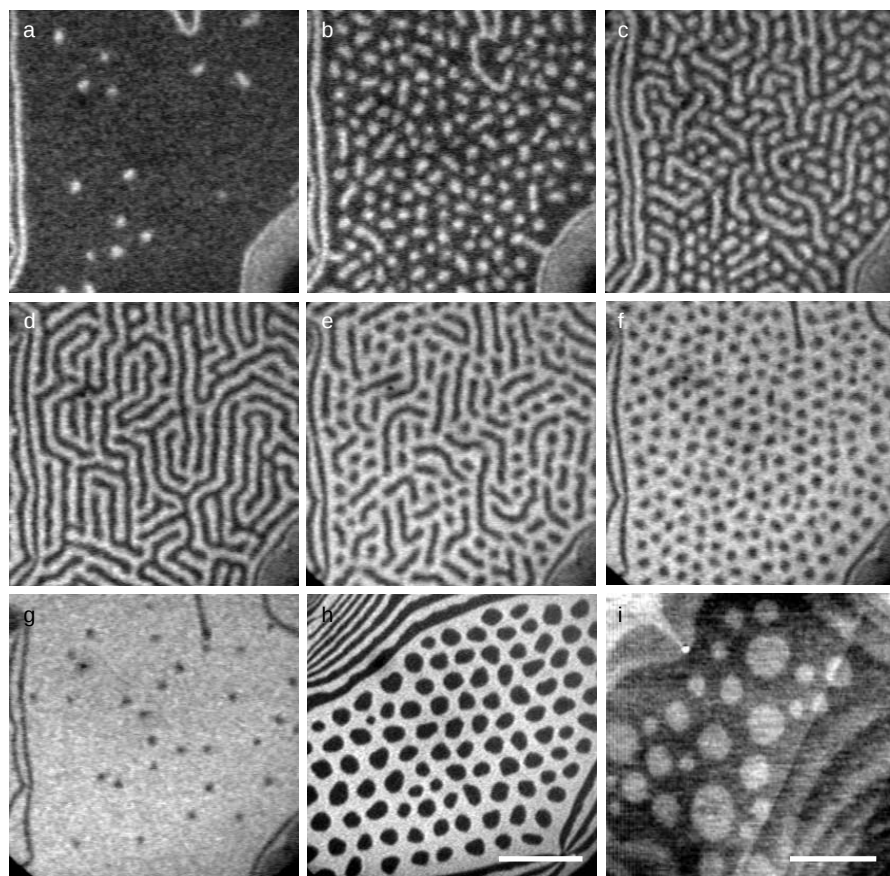


Figure 1 Self-assembly of Pb on Cu(111). Low-energy electron micrographs of the Cu(111) surface at 673 K with different area fractions of the lead-overlayer phase (bright) in the surface alloy phase (dark). **a–g**, Area fractions 0.03, 0.28, 0.35, 0.50, 0.65, 0.73 and 0.95, respectively. The domain pattern evolves from circular islands (droplets) to stripes, to vacancy islands (inverted droplets) with increasing lead coverage. **h**, Ordered droplet configuration at 623 K. Scale bar, 0.5 μm . **i**, Atomic-force micrograph of a droplet pattern after cooling down to room temperature and 2 hours of exposure to air. Scale bar, 0.3 μm .

Malaria

Cooperative silencing elements in var genes

Each *Plasmodium falciparum* malaria parasite carries about 50 *var* genes from a diverse family that encode variable adhesion proteins on the infected red blood cells of the host^{1–3}, but individual parasites single out just one *var* gene for expression and silence all the others^{4,5}. Here we show that this silencing is established during the DNA-synthesis phase (S phase) of the cell cycle and that it depends on the

cooperative interaction between two elements in separate control regions of each *var* gene (the 5'-flanking region and the intron). This finding should help to clarify the mechanisms by which parasites coordinate the silencing and activation of *var* genes that are responsible for antigenic variation in malaria.

The proteins (PfEMP1) encoded by the *var* genes enable red blood cells infected with *P. falciparum* to adhere within the circulatory system, thereby avoiding destruction by the spleen. Different PfEMP1 forms can attach to different host molecules,

leading to patterns of sequestration that produce disease complications such as cerebral malaria or the severe malaria of pregnancy^{6,7}. Because the exposed PfEMP1 molecules are subject to antibody attack during infection, *P. falciparum* parasites produce subpopulations with antigenically diverse PfEMP1 forms which must be continually chased by the immune response. The underlying switches in *var*-gene expression that result in this process of antigenic variation have yet to be understood.

Silent *var* promoters become transcriptionally active when removed from their chromosomal context⁸, suggesting that further control elements in the intact *var* gene are required to control or silence expression. Possible locations for such control elements are in the conserved *var* introns, which separate the two exons of all *var* genes and have sequences that are found in many heterogeneous sterile transcripts in erythrocyte-stage parasites².

To test whether *var* introns function in gene control, we inserted a representative intron sequence (*int*) into two different plasmids containing a luciferase (*luc*) reporter driven either by a *var* promoter sequence or by a promoter sequence from the unrelated gene *hrp3*, which encodes histidine-rich protein 3 (pVLH and pHLH, respectively; Fig. 1a). Transfection experiments with pVLH and pVLH/*int* indicated that the presence of *int* downstream of *luc* strongly repressed the luciferase signal (Fig. 1b, left), irrespective of *int* orientation. In contrast, the presence of *int* in the pHLH/*int* plasmid caused no detectable change in the reporter activity (Fig. 1b, right). The fact that the *var*, but not the *hrp3*, promoter was affected indicates that *int* repression is specific and occurs by cooperative action with the upstream *var* 5'-flanking region.

These initial experiments revealed strong repression, but not complete silencing, of *luc* expression from pVLH/*int* (Fig. 1b). Transition through S phase of the cell cycle is known to be required for the assembly of silent chromatin structures in other organisms⁹. Because parasites continually take up DNA during their growth in plasmid-loaded erythrocytes¹⁰, our experiments may have included heterogeneous populations of plasmids acquired before and after S phase, leading to incomplete silencing of the *var* promoter.

To test this possibility, we measured *luc* expression from homogeneous plasmid populations before and after transition through S phase. Assays done 18 h after invasion of plasmid-loaded erythrocytes

(before the onset of S phase) confirmed that *luc* was expressed without repression (Fig. 1c, left). After their maturation to mature schizont stages, we separated transfected parasites from plasmid-loaded erythrocytes using Percoll/sorbitol gradients. Returning these parasites to fresh cultures with plasmid-free erythrocytes eliminated further plasmid uptake after re-invasion and ensured that all plasmid DNA in the parasites had been through S phase. Complete reporter silencing from pVLH/*int* was then evident (Fig. 1c, right). We detected no silencing in control transfections with the original pVLH, or with a pVLH/*hsp* construct in which the *int* sequence was replaced with a *P. falciparum* *hsp86* intron of similar size (Fig. 1c, right).

The regulatory elements and S-phase-dependent silencing we identify here implicate cooperative DNA-binding complexes and modifications of chromatin structure in the control of *var*-gene transcription. Experimentally manipulable plasmid systems that incorporate mechanisms for *var* control may offer a useful approach by which to identify and characterize these complexes.

Kirk W. Deitsch, Michael S. Calderwood, Thomas E. Wellems

Laboratory of Parasitic Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892-0425, USA
e-mail: tew@helix.nih.gov

1. Baruch, D. I. et al. *Cell* **82**, 77–87 (1995).
2. Su, X.-Z. et al. *Cell* **82**, 89–100 (1995).
3. Smith, J. D. et al. *Cell* **82**, 101–110 (1995).
4. Chen, Q. et al. *Nature* **394**, 392–395 (1998).
5. Scherf, A. et al. *EMBO J.* **17**, 5418–5426 (1998).
6. Newbold, C. I. *Curr. Opin. Microbiol.* **2**, 420–425 (1999).
7. Scherf, A., Pouvelle, B., Buffet, P. A. & Gysin, J. *Cell Microbiol.* **3**, 125–131 (2001).
8. Deitsch, K. W., del Pinal, A. & Wellems, T. E. *Mol. Biochem. Parasitol.* **101**, 107–116 (1999).
9. Firestein, R., Cui, X., Huie, P. & Cleary, M. L. *Mol. Cell Biol.* **20**, 4900–4909 (2000).
10. Deitsch, K. W., Driskill, C. L. & Wellems, T. E. *Nucleic Acids Res.* **29**, 850–853 (2001).

correction

An unusual social display in gorillas

R. J. Parnell, H. M. Buchanan-Smith
Nature **412**, 294 (2001)

Our claim that the manipulation of water for communicative means by western lowland gorillas had not previously been described in any wild primate was incorrect: we were unaware of observations of a small number of male chimpanzees in a community in Tanzania throwing stones into streams¹, apparently for the purpose of intimidation.

1. Nishida, T. in *The Great Ape Project: Equality Beyond Humanity* (eds Cavalieri, P. & Singer, P.) 24–26 (Fourth Estate, London, 1993).

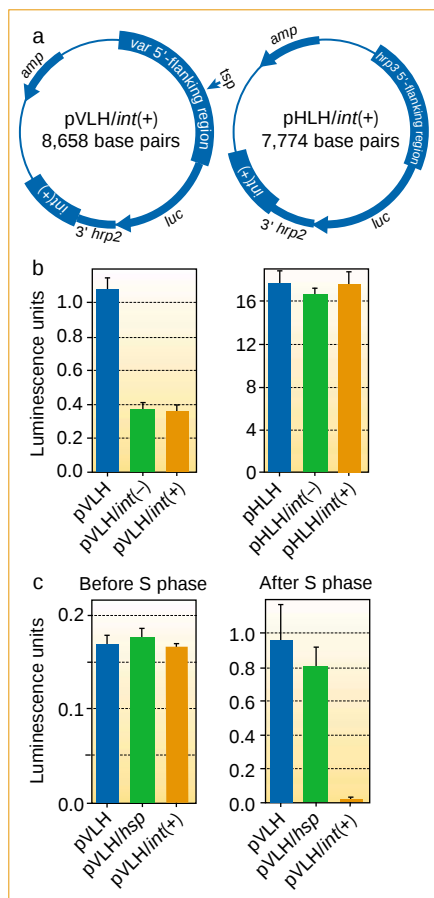


Figure 1 The role of cooperative control elements in S-phase-dependent silencing of *var* expression. **a**, Maps of pVLH/*int*(+) and pHLH/*int*(+). Plasmids pVLH/*int*(+), pVLH/*int*(-), pHLH/*int*(+) and pHLH/*int*(-) were constructed from the pVLH (pVLH-1) and pHLH (pHLH-1) plasmids⁸ by inserting a *var* intron (*int*) in the (+) or (-) orientation downstream of the 3'-flanking region of the *hrp2* gene. Plasmid pVLH/*hsp* was obtained by inserting the intron sequence from the unrelated *P. falciparum* *hsp86* gene into pVLH at the same position. GenBank accession numbers for pVLH/*int*(+), pVLH/*hsp* and pHLH/*int*(+) sequences are AF338824, AF379854 and AF338825, respectively. The transcription start point (tsp) for *var* is indicated. **b**, Luciferase activity from *P. falciparum* parasites after spontaneous transformation with original and modified pVLH and pHLH plasmids. Assays were carried out after cultivation of parasites for 72 h in the presence of plasmid-loaded erythrocytes¹⁰. **c**, Luciferase activity from homogeneous plasmid populations before and after S phase in synchronized transfected parasites. Experiments were carried out in triplicate and repeated on at least three separate occasions. Error bars indicate calculated standard deviations.

brieF communications is intended to provide a forum for both brief, topical reports of general scientific interest and technical discussion of recently published material of particular interest to non-specialist readers. Priority will be given to contributions that have fewer than 500 words, 10 references and only one figure. Detailed guidelines are available on *Nature's* website (www.nature.com) or on request from nature@nature.com

Identification of an angiogenic mitogen selective for endocrine gland endothelium

Jennifer LeCouter*, Joe Kowalski*, Jessica Foster†, Phil Hass‡, Zemin Zhang†, Lisa Dillard-Telm§, Gretchen Frantz§, Linda Rangell||, Leo DeGuzman*, Gilbert-Andre Keller||, Franklin Peale§, Austin Gurney†, Kenneth J. Hillan§ & Napoleone Ferrara*

Departments of * Molecular Oncology, † Molecular Biology, ‡ Protein Chemistry, § Pathology and || Toxicology/Pathology, Genentech Inc., South San Francisco, California 94080, USA

The known endothelial mitogens stimulate growth of vascular endothelial cells without regard to their tissue of origin. Here we report a growth factor that is expressed largely in one type of tissue and acts selectively on one type of endothelium. This molecule, called endocrine-gland-derived vascular endothelial growth factor (EG-VEGF), induced proliferation, migration and fenestration (the formation of membrane discontinuities) in capillary endothelial cells derived from endocrine glands. However, EG-VEGF had little or no effect on a variety of other endothelial and non-endothelial cell types tested. Similar to VEGF, EG-VEGF possesses a HIF-1 binding site, and its expression is induced by hypoxia. Both EG-VEGF and VEGF resulted in extensive angiogenesis and cyst formation when delivered in the ovary. However, unlike VEGF, EG-VEGF failed to promote angiogenesis in the cornea or skeletal muscle. Expression of human EG-VEGF messenger RNA is restricted to the steroidogenic glands, ovary, testis, adrenal and placenta and is often complementary to the expression of VEGF, suggesting that these molecules function in a coordinated manner. EG-VEGF is an example of a class of highly specific mitogens that act to regulate proliferation and differentiation of the vascular endothelium in a tissue-specific manner.

Embryonic development, reproductive functions, tumorigenesis and many other proliferative processes are critically dependent on the development of a vascular supply¹. Over the last decade, much progress has been made in the identification of the regulators of blood vessel growth. Most notably, VEGF and the Tie2 ligands are widely distributed and have been shown to play a coordinated role in endothelial cell proliferation and in the assembly of the vessel wall in a variety of normal and abnormal circumstances^{2–4}.

Evidence also exists for local, tissue-specific regulation of endothelial cell growth and differentiation^{5–7}. Endothelial cells of different organs exhibit a striking degree of morphological and functional diversity to adapt to local needs⁸. Whereas endothelial lineage is genetically determined, expression of a set of vascular-bed-specific genes is regulated at the transcriptional level after interaction with local signalling pathways⁶. The nature of the local regulatory signals is virtually unknown, although the specialized stroma is believed to be a major source⁷. Conceivably, an integrated network of stimuli, which may include tissue-specific secreted proteins in addition to cellular and extracellular matrix components, functions to determine the structure and function as well as modulate growth of the resident endothelium.

The microvascular endothelium of endocrine glands is highly permeable to blood-borne substances and to specific secretory products, reflecting the need for rapid hormone release in responses to homeostatic changes⁸. A major aspect of the high permeability in endocrine gland endothelium, as well as in some other organs, is the specialized discontinuities of the plasma membrane called fenestrae⁹. On the basis that VEGF mRNA is highly expressed around fenestrated microvessels, it has been proposed that VEGF is uniquely involved in the induction and maintenance of such structures¹⁰. However, expression of VEGF mRNA is also high in areas where capillaries are continuous (for example, myocardium or lung alveoli)¹¹ and therefore it is unlikely that VEGF is the only determinant of a fenestrated phenotype.

Here we report on the identification of an angiogenic mitogen

selective for one endothelial cell type, endocrine gland endothelium. Remarkably, the expression of this factor is largely restricted to steroidogenic glands. Although this protein does not show any structural homology to the VEGF family, it displays several striking biological similarities to VEGF, including hypoxic regulation and the ability to induce fenestration in the target cells. On the basis of such functional homologies, we propose the designation of endocrine-gland-derived vascular endothelial growth factor (EG-VEGF) for this factor.

Identification and structure of EG-VEGF

A library of purified human secreted proteins was screened for the ability to induce proliferation in primary bovine adrenal-cortex-derived capillary endothelial (ACE) cells. These cells respond only to a very restricted number of mitogens, including basic fibroblast growth factor (bFGF) and VEGF¹². We found that one factor, referred to herein as EG-VEGF, was capable of inducing a strong and reproducible mitogenic response. Mature EG-VEGF is a protein with a relative molecular mass of 8,600 (M_r 8.6K) encoded by a complementary DNA cloned from a human ovary library. The 1.4-kilobase (kb) cDNA encodes a protein of 105 amino acids, with a well defined signal sequence. The mature protein is predicted to have 86 amino acids, including 10 cysteines, and an expected isoelectric point (pI) of 8.46 (Fig. 1a). EG-VEGF displays a high degree of homology (80%) to a non-toxic protein purified from the venom of the black mamba snake (*Dendroaspis polylepis polylepis*), venom protein A (VPRA)¹³ (Fig. 1b). The structure of the native VPRA was solved and the disulphide bridge partners were revealed on the basis of a floating assignment by nuclear magnetic resonance analysis¹⁴. Notably, the number and spacing of cysteines for EG-VEGF are completely conserved. Thus, EG-VEGF is the human orthologue or a closely related homologue of VPRA. A human molecule closely related to a peptide isolated from the yellow-bellied toad (*Bombina variegata*), Bv8 (ref. 15), is 58% identical to the EG-VEGF mature protein and thus belongs to the same gene family. EG-VEGF displays

a more limited but significant homology with the cysteine-rich carboxyl sequence of the *Xenopus* head-organizer protein dickkopf, an inhibitor of wnt signalling¹⁶, and to colipase¹⁷ (Fig. 1c).

Angiogenic factors, including bFGF and VEGF, interact with extracellular matrix components, such as heparin sulphate proteoglycans, and this interaction is thought to regulate the bioavailability and activity of these molecules¹⁸. We tested whether EG-VEGF binds to heparin. When applied to a heparin sepharose column, EG-VEGF eluted in the presence of ~0.5 M NaCl, whereas the 165-amino-acid isoform of VEGF (VEGF₁₆₅) eluted at ~0.7 M NaCl under the same chromatographic conditions. This indicates that EG-VEGF is a heparin-binding growth factor and thus it may be sequestered in the extracellular compartment *in vivo*.

EG-VEGF is mitogenic and chemoattractive

EG-VEGF stimulated proliferation of ACE cells with a ED₅₀ (effective dose for a 50% response) of 0.2 nM and a maximal effect at ~2 nM (Fig. 2a). The fold increase in cell number was very similar to that induced by VEGF. The baculovirus-produced His- and Fc-tagged EG-VEGF proteins behaved very similarly in all experiments. Consistent with a mitogenic function, EG-VEGF induces in ACE cells a rapid and significant phosphorylation of the MAP kinases ERK1 and ERK2, as well as of other proliferation and survival signalling molecules (manuscript in preparation). The mitogenic activity of EG-VEGF was not blocked by administration of a soluble VEGF receptor (mFlt-immunoglobulin-γ (IgG)) tested at 50–1,000 ng ml⁻¹, indicating that such an effect is not mediated by VEGF release. These findings however do not rule out the possibility of reciprocal inductive effects between the two factors in more complex systems *in vivo*.

To determine the spectrum of cell types that proliferate in response to EG-VEGF, we tested a variety of other endothelial and non-endothelial cell types, including human umbilical vein (HUVEC), human dermal microvascular (HMVEC), bovine brain capillary (BBC) and adult bovine aortic (ABAE) endothelial cells. BBC cells showed about 10% of the proliferative response to VEGF, and the other endothelial cell types failed to show any effect after EG-VEGF administration, at all concentrations tested (Fig. 2b). EG-VEGF did not elicit any proliferative response in cultures of vascular smooth muscle cells (VSMCs), pericytes, fibroblasts or keratinocytes (Fig. 2c). Thus, EG-VEGF is a mitogen not only specific to endothelial cells, but also acting selectively on a defined type of endothelial cell. In contrast, other endothelial cell mitogens, such as VEGF and bFGF, do not show any significant selectivity for various endothelial cell types¹².

An essential aspect of the angiogenesis cascade is chemotaxis¹⁹. VEGF and bFGF are able to act as chemoattractants and stimulate endothelial cell migration. To further define the biological activity of EG-VEGF, we assessed whether ACE cells display a chemotactic response to this molecule in a modified Boyden chamber assay. A strong and reproducible chemotactic response was elicited in ACE cultures, with a peak response at 0.5 nM EG-VEGF (Fig. 2d). The magnitude of the effect was in the same range as that induced by VEGF.

To extend our results to another primary cell type, we purified endothelial cells from baboon adrenal cortex (BAEC). In BAEC cultures, 0.5–5 nM EG-VEGF induced a chemotactic response of the same magnitude as that induced by VEGF. Furthermore, EG-VEGF induced migration in MS-1 endothelial cells to an even greater extent than VEGF. The MS-1 cell line, isolated from microvessels of murine endocrine pancreas, retains highly differentiated properties, such as VEGF receptor expression²⁰. However, no response was elicited in HUVECs, even though these cells displayed a strong response to VEGF. Therefore, in addition to its mitogenic activity, EG-VEGF also acts as a chemoattractant, but only for a specific endothelial cell type. In agreement with these findings, binding studies with iodinated EG-VEGF revealed high-affinity, specific binding in ACE cells, which could be crosslinked to a membrane protein. No high-affinity binding was, however, detected in non-responsive cells such as HUVECs (manuscript in preparation).

EG-VEGF and VEGF induce fenestration

Endothelial cells within the endocrine glands display a rather unique fenestrated phenotype that is found in other restricted sites, such as kidney glomeruli, choroid plexus, median eminence, liver and many tumours⁸. Fenestrae are specialized plasma membrane microdomains, or windows, about 60 nm in diameter, that are usually clustered⁹. They are highly permeable to fluid and small solutes and are thought to facilitate large exchange of materials between interstitial fluid and plasma, such as that occurring in steroid-producing endocrine glands and others such as pancreatic

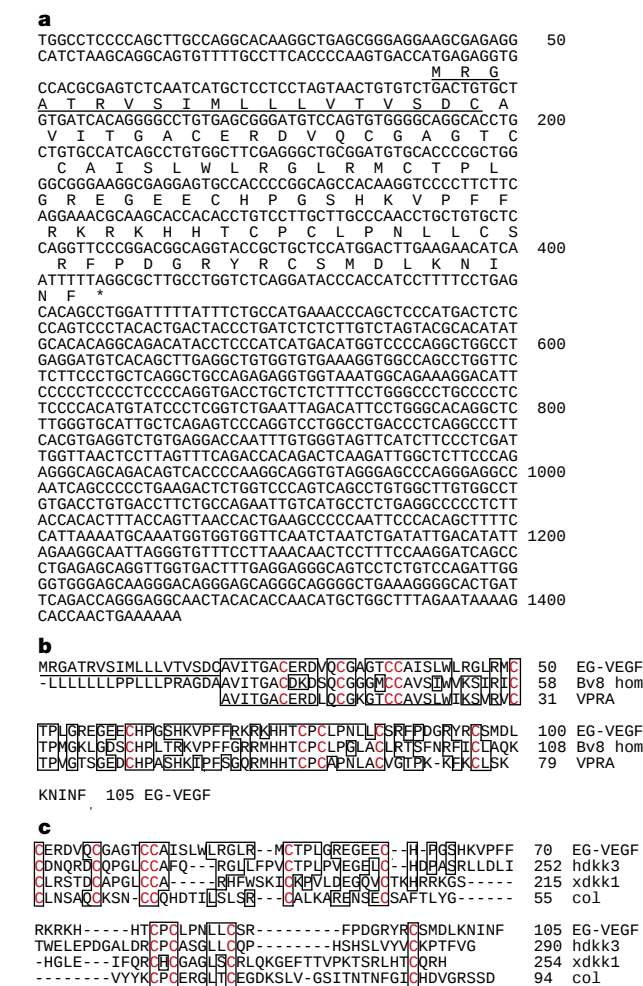


Figure 1 EG-VEGF cDNA and amino-acid sequences and alignments with homologous proteins. **a**, The 1.4-kb human EG-VEGF cDNA encodes a protein of 105 amino acids, with a classical signal sequence of 19 amino acids and a mature protein of 86 amino acids. A striking feature of the primary protein structure is the cysteine content, ten residues that potentially form five disulphide bridges. **b**, Human Bv8 homologue (Bv8 hom) shares 76% homology and 58% identity and the black mamba VPRA shares 80% homology and 63% identity with human EG-VEGF, and the cysteines are completely conserved among these proteins. Boxed residues indicate identity; human EG-VEGF, human Bv8 homologue and snake VPRA amino-acid sequences are indicated. **c**, Alignment of human EG-VEGF, human dickkopf-3 (hdkk3)⁴⁵, *Xenopus* dickkopf-1 (xdkk1)¹⁶ and porcine colipase (col) illustrates the conserved cysteines that form the characteristic disulphide-bonding pattern of this protein domain—the colipase fold⁴⁶. This motif in EG-VEGF is 37% identical and 41% homologous to the cysteine-rich C-terminal domain of human dickkopf-3; and 32% identical, 42% homologous to the *Xenopus* dickkopf-1 domain. Numbers indicate amino-acid position in the respective protein, and boxed residues are identical to EG-VEGF.

islets. We tested whether EG-VEGF could induce fenestration in endothelial cells, alone or in combination with VEGF. Notably, so far only VEGF has been reported to induce fenestration, *in vivo*²¹ and *in vitro*²².

ACE cells were grown to confluence on extracellular matrix produced by bovine corneal endothelial cells²³ and then treated with growth factors for 24 h. In agreement with previous studies²², VEGF induced fenestration in $4.33 \pm 1.53\%$ (mean $\pm$ s.d.) of ACE cell profiles (Fig. 2f). EG-VEGF had a very similar effect, inducing fenestration in $5 \pm 1\%$ of cell profiles (Fig. 2g). A combination of the two factors produced an additive or cooperative response, inducing fenestration in $11 \pm 1\%$ of the ACE cell profiles. No fenestrae were observed in the absence of VEGF or EG-VEGF, or in the presence of EGF. Very similar results were obtained with the islet-microvessel-derived MS-1 cell line (data not shown). These findings support the hypothesis that VEGF and EG-VEGF may cooperate *in vivo*, in settings such as the adrenal cortex or the ovary to induce the fenestrated phenotype and regulate permeability of the resident endothelial cells. It remains to be established whether EG-VEGF has any effects on fenestrated endothelia outside endocrine glands.

EG-VEGF expression is induced by hypoxia

Hypoxia is a key inducer of angiogenesis in both physiological and

pathological conditions^{24,25}. By virtue of activating hypoxia-inducible factor-1 (HIF-1), a dimeric transcription factor composed of HIF-1 α and HIF-1 β subunits, low oxygen tension is known to induce expression of VEGF and Flt-1 (VEGFR-1), in addition to erythropoietin (EPO) and certain glycolytic enzymes, through *cis*-acting regulatory elements²⁴. We sought therefore to determine whether hypoxia regulates expression of EG-VEGF mRNA. Exposure of the human adrenal carcinoma cell lines SW13 and H295R to hypoxic conditions ($\sim 2\%$ oxygen) resulted in a $275 \pm 15\%$ and $210 \pm 12\%$ increase in EG-VEGF mRNA levels above normal, respectively, whereas the VEGF mRNA increased $352 \pm 30\%$ and $266 \pm 15\%$, respectively (Fig. 3a). Similar results were obtained in two additional independent experiments.

A database search of the *EG-VEGF* promoter sequence for the core HIF-1 binding site revealed a putative element within the first 2,450 nucleotides upstream of the transcription start site (Fig. 3b), based on a consensus sequence (TACGTGCGGC, bold represents the invariable sequence). After an 18–24-h incubation, a luciferase reporter construct containing the putative *EG-VEGF* element conferred a 3.3 ± 0.8 -fold increase above normal oxygen conditions. This level is comparable to that conferred by the hypoxia-response element (HRE) *Epo* consensus, 3.4 ± 1.2 -fold (Fig. 3c). Mutating the core sequence of either the consensus or the putative *EG-VEGF* HRE abolished the response to hypoxia, verifying the specificity of

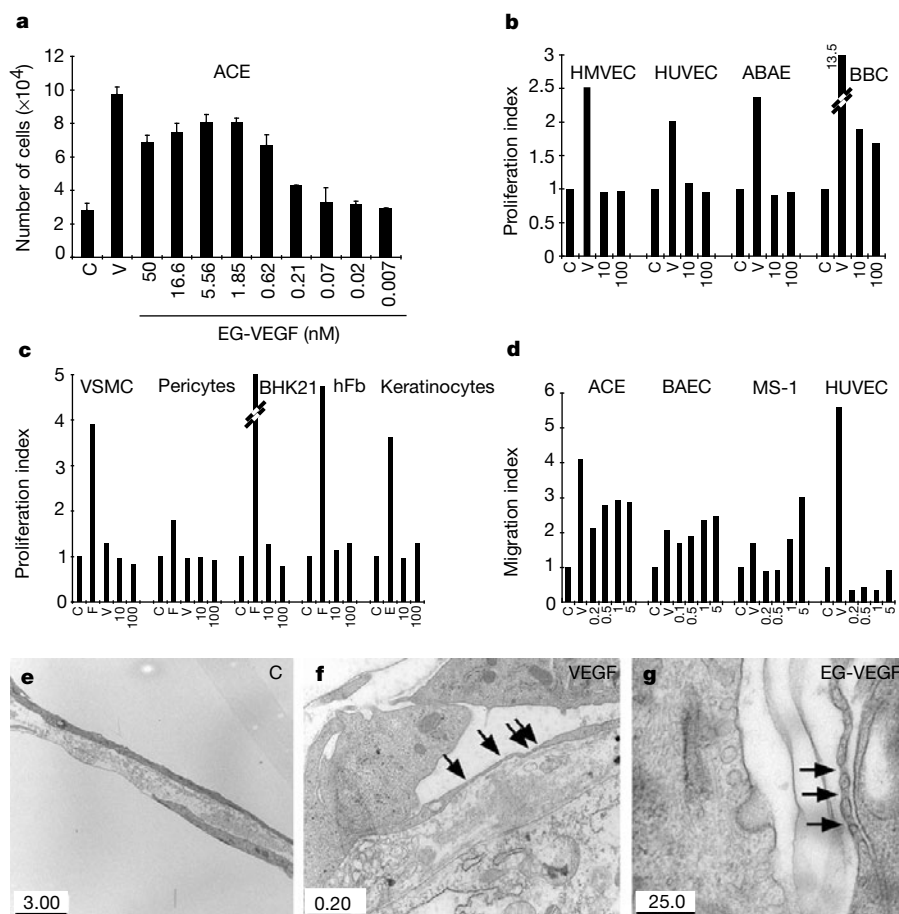


Figure 2 EG-VEGF is a mitogen, chemoattractant, and inducer of fenestration for specific endothelial cells. **a**, In ACE cell cultures, EG-VEGF induced a maximal mitogenic response at 2 nM, with a ED_{50} of 0.2 nM. C, control. **b**, **c**, Proliferation assays with several endothelial (**b**, HUVECs, HMVECs, BBC cells, ABAE cells) and non-endothelial cell types (**c**, VSMCs, pericytes and fibroblasts (BHK21 and human neonatal fibroblasts, hFb) and keratinocytes). Basal media served as a negative control (C); bFGF (F), EGF (E) or VEGF (V), added respectively at 5, 5 and 10 ng ml⁻¹, served as positive controls. EG-VEGF was tested at 10 and 100 nM. **d**, EG-VEGF induced a chemotactic response in ACE cells,

BAECs and MS-1 cells, but not in HUVECs. Each graph is a representative experiment. Data are mean $\pm$ s.d., and the proliferation or migration index is relative to the negative control, arbitrarily set to 1. **e–g**, EG-VEGF induces fenestration in adrenal cortex-derived capillary endothelial cells. ACE cells, grown to confluence on ECM, were treated with 2.5 nM VEGF, 10 nM EG-VEGF or both. Electron micrographs of ACE cells untreated (**e**), treated with VEGF (**f**) or EG-VEGF (**g**) revealed that both molecules are capable of inducing fenestration. Arrows indicate the location of fenestrae. Magnification is indicated (in μ m for **e** and **f**; nm for **g**).

the response. Although we cannot rule out additional mechanisms, pending a detailed promoter analysis these findings indicate that in all likelihood HIF-1 is a mediator of the hypoxic regulation of the EG-VEGF gene. It is striking that, in spite of the lack of homology, EG-VEGF and VEGF not only have similar effects in responsive endothelial cells but may also be regulated through common mechanisms.

EG-VEGF is restricted to steroidogenic tissues

To elucidate the expression pattern of EG-VEGF, we performed dot blot analysis on tissue RNA arrays representing an extensive series of human tissues and cell lines. This evidenced expression only in ovary, testis, adrenal, placenta and, to a lesser extent, prostate (see Supplementary Information). To verify this finding, we performed

northern blot analysis using RNA from a variety of human tissues. A single mRNA species of 1.4 kb was expressed, in decreasing order of intensity, in ovary, testis, adrenal and placenta, in agreement with the dot blot data. No expression was detected in any other tissue examined, with the exception of a very weak signal in the prostate (Fig. 4). Similar findings were obtained by *in situ* hybridization in human multi-tissue arrays (data not shown). These findings indicate that steroidogenic endocrine glands are the major site of expression of EG-VEGF mRNA. A much weaker expression occurs in a non-steroidogenic organ like the prostate.

To identify the cell types that express EG-VEGF, we examined a series of ovary and testis specimens from humans and other primates. In the testis, we detected a strong hybridization signal, which was virtually restricted to the Leydig cells, the testosterone-producing cells (Fig. 5a–d). It is noteworthy that the endothelium of the testis has a surprisingly high turnover: as many as 3% of endothelial cells are labelled with bromodeoxyuridine (BrdU), which is probably related to the intense metabolic activity associated with spermatogenesis and steroidogenesis²⁶. In this context, Leydig cells are known to be a source of angiogenic and permeability-enhancing factors such as VEGF²⁶. Notably, in our specimens the VEGF hybridization signal was considerably less intense than the EG-VEGF signal (data not shown). In the ovary, intense EG-VEGF mRNA expression was localized to the cortical stroma, the cumulus oophorus, theca and granulosa of developing follicles (Fig. 5e–p). The strong expression of EG-VEGF mRNA within the theca is coincident with the development of a capillary network associated with follicular development and steroid hormone production²⁷, and is consistent with a pro-angiogenic role of EG-VEGF.

This expression pattern is essentially similar in the other primate species examined, the cynomolgus monkey (*Macaca fascicularis*) and the chimpanzee (*Pan troglodytes*) (Fig. 5g–j and data not shown). Whereas VEGF has a low expression in the specialized stroma²⁸, EG-VEGF is strongly expressed (Fig. 5e, f, k, l). Interestingly, such stroma becomes intensely hyperplastic and angiogenic, producing excessive amounts of androgens, in the polycystic ovary syndrome²⁹, and we observed strong expression of EG-VEGF mRNA in the hyperplastic stroma in a series of human polycystic ovary specimens (unpublished observations). The expression pattern of EG-VEGF in the cumulus oophorus is very similar to that of VEGF^{28,30} (Fig. 5m–p). EG-VEGF expression is not detected in the corpus luteum (data not shown). Whereas EG-VEGF is highly expressed in the ovarian stroma and within developing follicles, VEGF is expressed at the highest level in the corpus luteum^{28,30} (Fig. 5e–p and data not shown).

EG-VEGF induces angiogenesis and ovarian cyst formation

We examined the *in vivo* activity of EG-VEGF in a variety of model systems. In the rat corneal pocket assay, purified EG-VEGF failed to show a significant response in the eyes tested, whereas VEGF induced the expected angiogenic effects (Fig. 6a–c). Similar results were obtained in the rabbit cornea (data not shown). To achieve EG-VEGF delivery in a local and sustained fashion, we generated adenovirus vectors and injected these into various sites in athymic nude rats or mice. Recombinant adenovirus expressing lacZ or VEGF served as controls. When injected into the skeletal muscle of nude rats, the effects of lacZ and EG-VEGF adenovirus were essentially indistinguishable, with minimal inflammatory infiltrate and absence of angiogenesis (Fig. 6d–f). In contrast, the VEGF-expressing adenovirus induced frank angiogenesis, with mitotic endothelial cells, oedema and granulation tissue, in agreement with previous studies³¹. However, the EG-VEGF protein was produced at high level in the skeletal muscle, arguing against the possibility that lack of EG-VEGF synthesis may explain the absence of angiogenic effects in this site (see Supplementary Information). Very similar results were obtained after injection of the same adenovirus vectors in the mouse ear (Fig. 6g, h) or skeletal muscle (data not shown).

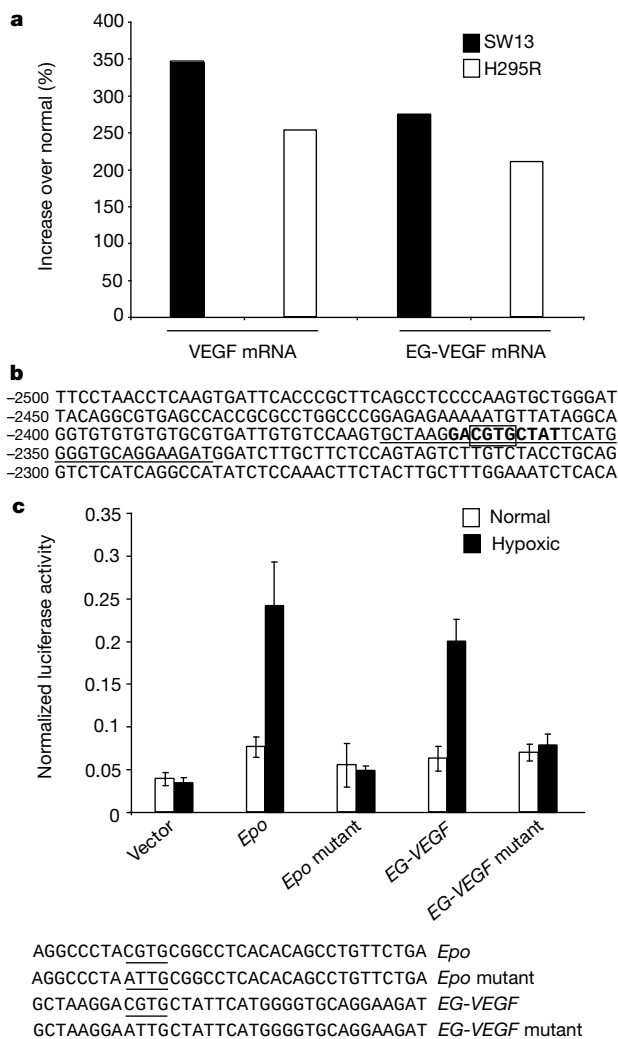


Figure 3 EG-VEGF expression is hypoxia inducible. **a**, Taqman analysis of RNA from SW13 and H295R human adrenal carcinoma cells exposed to normal-oxygen (22% O₂) or hypoxic (2% O₂) conditions for 18 h revealed that EG-VEGF transcription increased 275% and 210% in SW13 and H295R, respectively, comparable to the 350% and 252% increases in VEGF over the normal-oxygen controls. VEGF and EG-VEGF data were normalized to β -actin. **b**, A putative HRE (bold), with the invariable core boxed, is shown within the context of EG-VEGF 5' promoter sequence, 2,250–2,500 nucleotides upstream of the transcription start (GenBank accession number AC068519). **c**, HRE luciferase activities in normal-oxygen and hypoxic conditions. Activities of luciferase reporters and vector were normalized to the co-transfected *Renilla* luciferase. Both the *Epo* consensus and *EG-VEGF* constructs were induced ~3.4-fold by hypoxia above their respective normal-oxygen controls. Mutation of the core sequence, in *Epo mutant* and *EG-VEGF mutant*, abrogated the responsiveness to hypoxic conditions, resulting in activities similar to vector control.

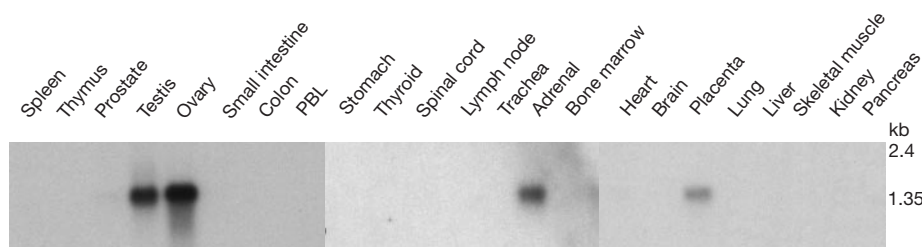


Figure 4 Northern blot analyses of human RNA samples from different tissues with an EG-VEGF specific probe. A single EG-VEGF transcript of ~1.4 kb was identified. Expression is highest in ovary and testis, followed by adrenal and placenta. A signal is detectable in prostate only after prolonged exposure. Equivalent RNA loading was assessed by

hybridization with the control actin probe (data not shown). Contents of the lanes are indicated above the blots, and the size is indicated at the right. PBL, peripheral blood leukocytes.

In contrast, intra-ovarian delivery of either EG-VEGF or VEGF adenovirus resulted within a week in a dramatic enlargement of the injected ovaries, with grossly visible blood vessels as well as haemorrhagic areas (Fig. 6i). The ovarian wet masses in a representative experiment in nude rats were 33 ± 3 mg in the lacZ group, 489 ± 30 mg in the EG-VEGF group and 191 ± 91 mg in the VEGF group ($n = 4$). The mass of uninjected ovaries (37 ± 7 mg) was not different from the lacZ group. Although there was a trend towards larger mass in the EG-VEGF group compared with the VEGF group, the histological picture was very similar. There was, in both cases, ovarian enlargement with disruption of the normal ovarian stroma by large cystic spaces containing proteinaceous material and frank haemorrhage (Fig. 6k, l). High-power examination revealed intense angiogenesis at the periphery of the cysts (Fig. 6m, n). The picture is consistent with cyst formation secondary to leakage from the newly formed vessels. The morphology of the lacZ ovaries was normal (Fig. 6j). Thus, VEGF and EG-VEGF are able to trigger the same

chain of events in a responsive tissue, including the ability to induce fluid extravasation, in addition to angiogenesis. This could have been expected, in view of the striking similarity of the two factors in their effect on responsive endothelial cells, including the ability to induce fenestration.

Discussion

VEGF was identified as a mitogen specific to endothelial cells more than ten years ago¹². Such selectivity, which distinguished it from other endothelial cell mitogens, led to the hypothesis that this molecule might play a unique role in the regulation of blood vessel growth¹². Subsequent studies indeed demonstrated the essential role of VEGF in a surprisingly broad variety of normal and abnormal angiogenic processes, including embryogenesis³², post-natal growth³³, corpus luteum development³⁴, growth plate morphogenesis³⁵, tumorigenesis³⁶ and ischemic retinal diseases³⁷. However, it was difficult to explain how a single, ubiquitously

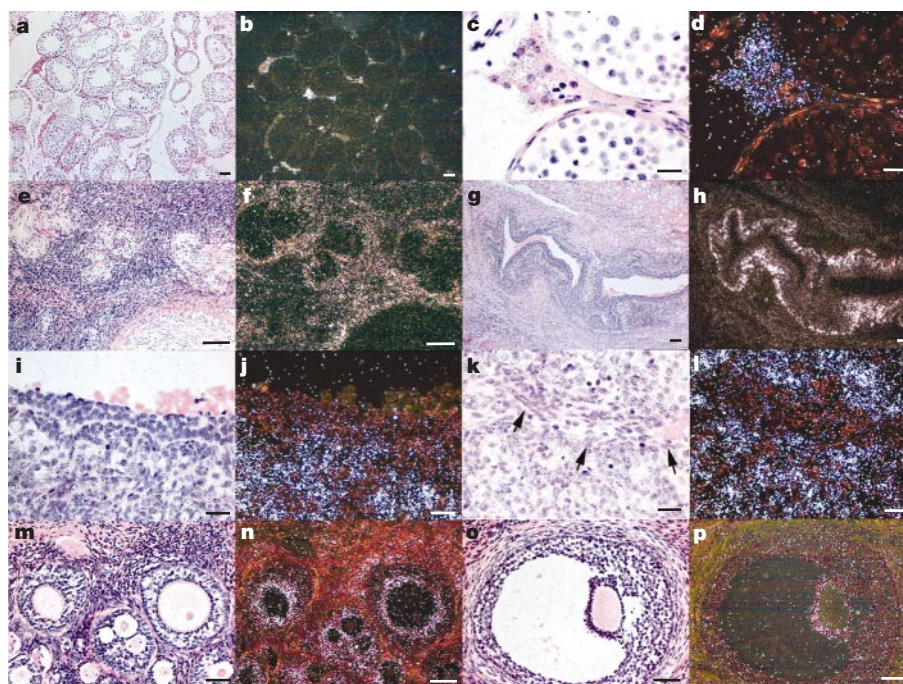


Figure 5 Expression of EG-VEGF mRNA in testis and ovary. Bright-field images and the corresponding dark-field images are shown (a and b, c and d, and so on). a–d, *In situ* hybridization studies revealed that, in the human testis, EG-VEGF transcript is restricted to the testosterone-producing Leydig cells. e, f, EG-VEGF signal is prevalent in the human ovarian stroma in cells that can be described as perivascular (k, l). A very similar pattern was evidenced in the chimpanzee ovary (data not shown). g, h, A strong signal is detected

in the human corpus haemorrhagicum. In developing follicles of humans (i, j) and chimpanzees (m–p), EG-VEGF RNA is expressed in the theca and granulosa cells—both steroidogenic cell types—and the cumulus oophorus. Arrows indicate the course of a blood vessel. Scale bars: a, b, g, h, 0.1 mm; c, d, i–l, 25 μ m; e, f, 115 μ m; m–p, 55 μ m.

expressed molecule might account for the great diversity of phenotypic and growth characteristics of the endothelia in those circumstances. Conceivably, while VEGF signalling represents a rate-limiting step in many situations, local regulatory pathways modulate endothelial cell growth and function in a tissue- or organ-specific manner, synergistically with VEGF. Furthermore, recent data suggest the existence of VEGF-independent angiogenic pathways that remain to be identified. For example, early results of clinical studies with VEGF inhibitors in cancer patients have

provided initial evidence of efficacy. However, progression eventually occurs in advanced malignancies, in spite of the anti-VEGF treatment³⁸.

Our results demonstrate that steroidogenic glands express a novel angiogenic and permeability-enhancing factor, EG-VEGF, with unique target cell selectivity. The partially complementary expression patterns of VEGF and EG-VEGF indicate that these molecules may function in a coordinated or complementary manner to regulate angiogenesis and permeability as well as to maintain a

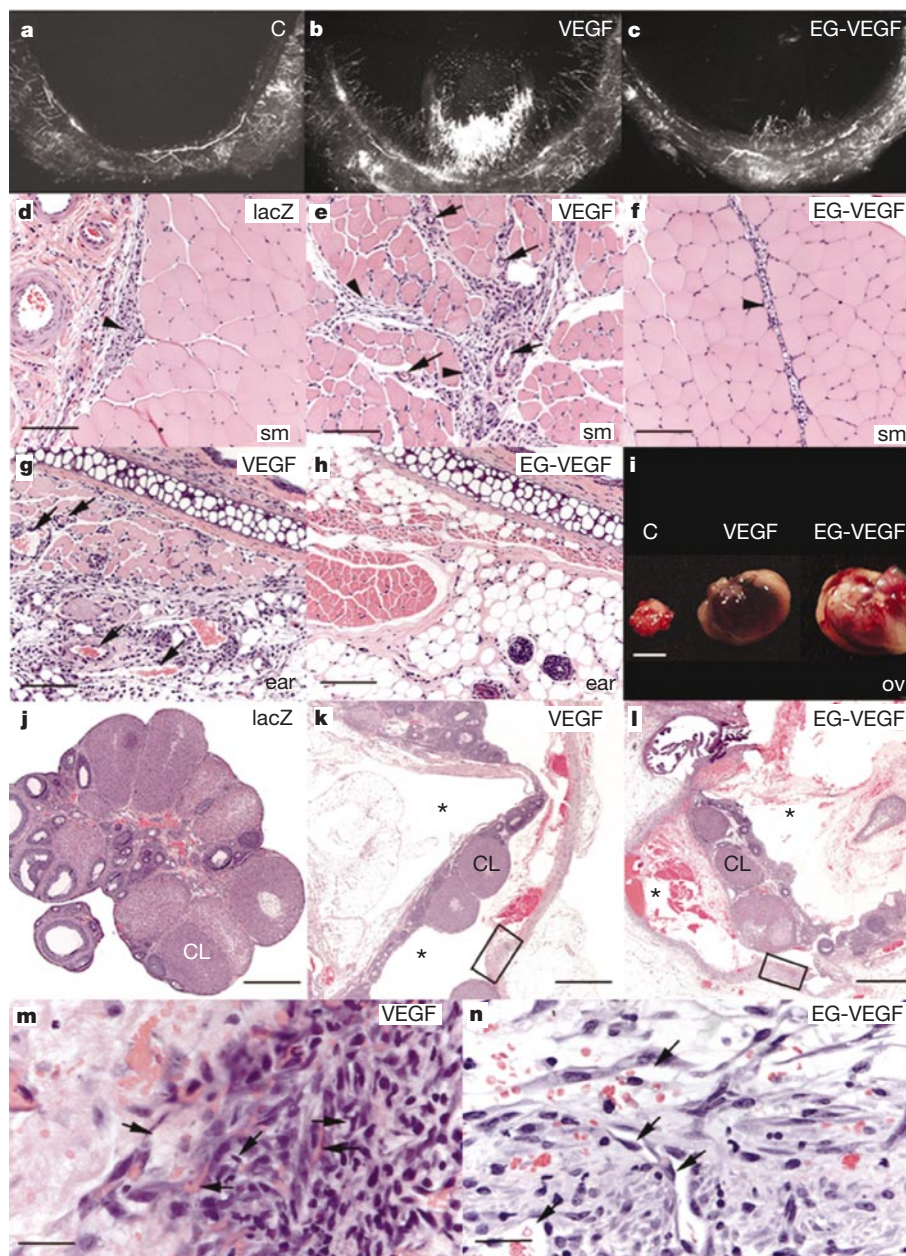


Figure 6 Selectivity of the *in vivo* angiogenic effects of EG-VEGF. **a–c**, Rat corneal pocket assay. Note the strong angiogenic response induced by VEGF protein, whereas EG-VEGF has essentially no effect. **d–f**, Injection of adenovirus CMV-lacZ, CMV-VEGF₁₆₄, or CMV-EG-VEGF (5×10^8 PFU) in the skeletal muscle (sm) of nude rats. Arrowheads indicate microspheres marking the injection site; arrows indicate new blood vessels. Note the angiogenic response, with abundant new vessel formation, induced by VEGF, whereas both lacZ and EG-VEGF adenovirus had no appreciable effects. **g, h**, VEGF and EG-VEGF adenovirus injection in the mouse ear. Again, VEGF adenovirus resulted in a strong angiogenic response, which was absent in the animals injected with EG-VEGF adenovirus. **i**, Gross appearance of ovaries (ov) following injection of lacZ (C, control), VEGF or

EG-VEGF adenovirus, after 7 d. Note the much larger mass plus the presence of superficial vessels and haemorrhagic areas in both VEGF and EG-VEGF groups. **j–l**, Micrographs of ovaries injected with adenovirus vectors (5×10^8 PFU), as indicated. Note the normal architecture and morphology of the lacZ-treated ovary in **j**. Corpora lutea (CL) are indicated. In contrast, the VEGF (**k**) and EG-VEGF (**l**) groups revealed very similar changes, with large fluid-filled or haemorrhagic cystic areas (asterisks). **m, n**, High-power micrographs of boxed areas in **k** and **l**, respectively. Areas of intense angiogenesis are evident at the periphery of cystic lesions. Arrows point towards blood vessels. Scale bars: **d–h**, 100 μ m; **i**, 0.5 cm; **j–l**, 1 mm; **m, n**, 33 μ m.

fenestrated phenotype in the ovary and in the other glands that co-express these factors. In this context, inactivation of VEGF in rodents³⁴ or in primates³⁹ results in a dramatic suppression of corpus luteum development. However, endothelial cell growth associated with follicular development is considerably less affected (unpublished observations), indicating that additional factors may contribute to follicular angiogenesis. EG-VEGF potentially contributes to such VEGF-independent angiogenesis. The study of EG-VEGF may offer insights into pathophysiology and therapy of endocrine disorders characterized by excessive angiogenesis, such as the polycystic ovary syndrome and perhaps also the ovarian hyperstimulation syndrome.

Hypoxia-induced angiogenesis is primarily an adaptive physiological response to an increase in metabolic demands associated with proliferative processes, including those occurring during the cyclical ovarian changes²⁵. That EG-VEGF is induced by hypoxia through a potential HIF-1-dependent mechanism indicates that this gene is in the same homeostatic cascade as VEGF, Flt-1 and erythropoietin, and lends further credence to the hypothesis that it has an important regulatory function. Regulation of EG-VEGF expression by hormones is potentially of major importance and remains to be investigated.

Further studies will be required to fully elucidate the spectrum of activity of EG-VEGF. The identification of the receptor(s) and analysis of their distribution will be critical steps. Even a selective molecule like VEGF has been found to have unexpected mitogenic and non-mitogenic activities in cell types other than endothelial cells, including neuronal cells⁴⁰. Interestingly, the snake protein VPRA has been reported to promote contractility in isolated ileum⁴¹. The significance of this activity remains to be established.

Our findings suggest a paradigm for the regulation of angiogenesis. Steroidogenic glands seem to have developed highly specific local mechanisms, possibly to 'fine tune' and complement the action of the ubiquitous VEGF/VEGF receptor and Tie2 systems. Perhaps this is not unique to endocrine glands and other regulatory proteins exist that act selectively on endothelial cells resident in various tissues and organs. For example, mitogens specific for the endothelium of heart or skeletal muscle could be of great therapeutic significance. A lack of activity in conventional assays and/or absence of homology to known families of endothelial cell mitogens may have so far precluded the discovery of such molecules. Further characterization of the molecular basis of the tissue-specific regulation of endothelial cell growth and differentiation should allow greater insight into the biology of the endothelium and refine our ability to inhibit or promote blood vessel growth.

Note added in proof: According to the most updated sequencing of human chromosome 1, the putative hypoxia-response element (HRE) identified at nucleotide -2,347 from the transcription start—based on GenBank accession number AC068519—is >10 kb upstream. Two additional potential HREs are present in the 5' untranslated region of the EG-VEGF gene, respectively at position -1,773 and -4,052 from the transcription start (GenBank accession number AL390797). □

Methods

Recombinant Fc/His fusion proteins

A cDNA clone encoding human EG-VEGF was identified in the Lifeseq EST database (Incyte Genomics) on the basis of an algorithm that predicted the signal peptide and was sequenced in its entirety. The coding sequences of human EG-VEGF were amplified by polymerase chain reaction (PCR) and subcloned into the *EcoRI* and *SmaI* sites of pBPH.IgG to generate a carboxy-terminal fusion with the Fc region of human IgG1 or into the *EcoRI* and *SmaI* sites of pBPH.His.c to generate a C-terminal GHHHHHHH tag. Vectors pBPH.IgG and pBPH.His.c are derivatives of the baculovirus expression vector pVL1393 (PharMingen).

The fusion proteins were expressed in H5 cells by recommended procedures (Invitrogen). In brief, the DNA constructs were cotransfected with BaculoGold baculovirus DNA (PharMingen) in a 7:1 ratio into adherent Sf9 cells. Cells were incubated at 28 °C for 2–4 d, and the supernatant was collected. The transfection supernatant was

amplified and subjected to affinity purification by either protein A sepharose beads (Pharmacia) for Fc fusion proteins or Ni-NTA agarose beads (Qiagen) for His-tagged proteins. To examine the protein expression, SDS-PAGE analysis was performed on the affinity-purified recombinant proteins under non-reducing and reducing conditions, followed by staining. Proteins were routinely submitted for amino-terminal sequencing and endotoxin measurements, which were below 1 endotoxin unit (EU) per mg. CHO cells expressing the untagged EG-VEGF were established and the highest-producing clones were expanded. To determine whether EG-VEGF binds to heparin, we applied ~15 µg of untagged protein into a 1-ml Hi-Trap heparin-sepharose column (Pharmacia) in 10 mM Tris, pH 7.4, 0.1 M NaCl. The column was eluted with a linear gradient of NaCl (0.1–2 M). Human VEGF₁₆₅ was used as a reference.

Proliferation assays

Media and other cell culture reagents were obtained from Life Technologies. ACE cell proliferation assays were performed as previously described¹². HUVECs, HMVECs and human keratinocytes were purchased from Clonetics. Bovine pericytes and human neonatal fibroblasts were a gift of M. Gerritsen. For proliferation assays, 4,000–6,000 cells per well were plated in 12-well dishes. Controls included wells in the basic assay media with no other added factors; bFGF or VEGF, in the case of endothelial cell assays, served as positive controls. Cell numbers were determined after 5–7 d.

Migration assays

ACE cells, BAECs, MS-1 cells and HUVECs were used for migration assays. Primary baboon endothelial cells were isolated from adrenal glands of premature or fetal baboons, which were a gift of R. Clyman. The capsule was removed and the remaining tissue was finely chopped with sterile razor blades into fragments ~2 mm³. The fragments were subsequently incubated at 37 °C for 30–40 min in 0.1% collagenase (Sigma) in 50:50 Ham's F10:DMEM media, 10% FCS with the addition of DNase I (Life Technologies). Single-cell suspensions were washed, re-suspended in PBS with 5% FCS, incubated with 1 µg of anti-KDR monoclonal antibody for 10⁷ cells, washed, and then incubated with goat anti-mouse antibody conjugated to fluorescein isothiocyanate (FITC; Sigma). The KDR-positive and -negative populations were sorted with the EPICS ELITE cell sorter (Coulter). The endothelial cell population was maintained in CSC medium containing 10% serum and growth factors (Cell Systems). Filter inserts (8.0 µm, Falcon 3097) were coated with 1% gelatin.

Cells were cultured before being assayed, as described above. Cells were digested with trypsin and transferred to endothelial basal media (EBM, Clonetics) with 0.1% BSA for the assay. Cells were plated at 1–5 × 10⁴ per upper chamber and growth factors were placed in the lower chamber. The assay was routinely 16 h at 37 °C. At completion, we removed cells from the upper side of the membrane by scraping with a polyurethane swab; the remaining cells on the bottom of the membrane were then fixed with methanol. Cells were stained with propidium iodide and counted under low power with the Image-Pro program.

Fenestration assays

Extracellular matrix (ECM) was prepared as previously described²³. Briefly, corneal endothelial cells were isolated from steer eyes (Pel Freez) and these were expanded in 50:50 Ham's F10:DMEM media supplemented with 15% FCS, penicillin-streptomycin, fungizone. To prepare the ECM-coated plates, 4 × 10⁴ cells were plated per well in 6-well dishes and cultured for ~10 d in low-glucose DMEM supplemented with 10% FCS, 2.5% dextran (Sigma) and penicillin-streptomycin. At the end of 10 d, the cells were quickly lysed in 0.02 M NH₄OH in water, rinsed several times with PBS and stored at 4 °C in PBS with antibiotics. ACE cells were plated at a density of 1–2 × 10⁵ and grown to confluence. No addition, 10 nM EGF, 2.5 nM VEGF₁₆₅, 10 nM EG-VEGF or 2.5 nM VEGF₁₆₅ plus 10 nM EG-VEGF were added to individual wells, in at least duplicate. The fenestration assays were replicated three times. Cells were rinsed with PBS and fixed for 2 h in 2% formaldehyde, 2.5% glutaraldehyde in 0.1 M cacodylate buffer. After washing, the samples were post-fixed in aqueous 1% osmium for 2 h, washed in water, dehydrated through graded ethanol and propylene oxide, and embedded in EPONATE 12 (Ted Pella). Ultrathin sections were cut on a Reichert Ultracut E microtome, counterstained with uranyl acetate and lead citrate and examined in a Philips CM12 transmission electron microscope at 80 kV; images were captured with a GATAN Retractable Multiscan digital camera. One hundred cell profiles were examined for each condition in triplicate.

Expression analyses

Human multiple tissue polyA⁺ RNA array and human polyA⁺ RNA multiple tissue northern blots were purchased from Clontech. cDNA probes were prepared using 30–50 ng of the human cDNA fragments with the Redi-Prime II kit (Amersham), using ³²P-dCTP at 3,000 µCi mmol⁻¹ (Amersham). Probes were purified on Sephadex G50 spin columns (Pharmacia) and hybridizations were carried out at 68 °C in ExpressHyb hybridization solution (Stratagene). Tissues were processed for *in situ* hybridization by a method described previously⁴². ³³P-UTP-labelled RNA probes were generated as described⁴³. Sense and antisense probes were synthesized from a cDNA fragment corresponding to nucleotides 219–958 of the human sequence.

Hypoxic regulation studies

For expression analysis, RNA isolates from replicate, matched samples of SW13 and H295R cells (both from the ATCC) treated with normal-oxygen or hypoxic conditions were prepared using the Rneasy kit (Qiagen) as described by the manufacturer. For real-time quantitative PCR with reverse transcription (RT-PCR), 50 ng of total RNA was

assayed in triplicate with the Perkin Elmer Taqman kit reagents and an ABI prism 7700 Sequence Detector. Oligonucleotides and probes used were as follows: EG-VEGF forward, 5'-CCGGCAGCCACAAGGTC; EG-VEGF reverse, 5'-TGGGCAAGCAAGGACAGG; EG-VEGF probe, FAM5'-CCTTCTTCAGGAAACGCAAGCACCAC-3'(TAMARA); VEGF forward, AATGACGAGGGCCTGGAGT; VEGF reverse, TTGATCCGCATATCTGCATG; VEGF probe, FAM5'-TGTGCCCACTGAGGAGTCCAACATCA-3'(TAMARA). β -actin primers and probe were obtained from the manufacturer (Perkin-Elmer).

The luciferase reporter constructs were generated by cloning compatible, annealed oligonucleotides into the BglII site of the pGL3-Promoter vector (Promega): HRE_consensus(Epo) sense, AGGCCCTACGTGCGGCCTCACACAGCCTGTTCTGA; HRE_mutant(Epo) sense, AGGCCCTAATTGCGGCCTCACACAGCCTGTTCTGA; HRE_EG-VEGF sense, GCTAAGGACGTGCTATTATCATGGGGTGCAGGAAGAT; HRE_EG-VEGF mutant sense, GCTAAGGAATTGCTATTATCATGGGGTGCAGGAAGAT. Transient transfections of HeLa cells were performed in 6-well dishes using Effectene transfection reagent (Qiagen). For each well, 0.75 μ g of reporter along with 0.25 μ g of pRL-SV40 control vector was transfected. Triplicate, paired transfections were used in parallel normal-oxygen versus hypoxic incubations, dishes being exposed to hypoxia for 18–24 h commencing 24 h after transfection. Cells were lysed and assayed using the dual-luciferase reporter assay system (Promega). Luciferase activity is normalized to control vector values, and a representative experiment is shown, error bars representing the s.d.

Adenovirus generation

Adenoviral CMV-hEG-VEGF, CMV-lacZ and mVEGF₁₆₄ were generated using the AdEasy adenoviral vector system (Stratagene) essentially as described by the manufacturer. The coding region of EG-VEGF or VEGF was cloned between the *NotI* and *HindIII* sites of the pShuttleCMV vector. These vectors, along with the supplied pShuttleCMV-lacZ, were recombined, in BJ5183 electrocompetent bacteria (Stratagene), with the AdEasy vector containing the Ad5 genome deleted for E1 and E3 regions. Primary viral stocks were prepared by transiently transfecting the recombined AdEasy plasmids into host HEK293 cells. Adenovirus stocks were further amplified in HEK293 cells and purified using the Virakit Adeno purification kit (Virapur). Adenovirus titres were obtained by agarose-overlaid plaque assays.

In vivo assays

Corneal pocket angiogenesis assays in adult rats were performed as previously described⁴⁴. Hydron-coated pellets containing 200 ng VEGF₁₆₅ or 500 ng EG-VEGF, plus methylcellulose and aluminium sucralphate, were inserted into the base of the pocket ($n = 6$). Control pellets contained 500 ng CD4-IgG. At day 6, animals were killed and injected with high-molecular-mass FITC-dextran to allow for visualization of the vasculature. Measurements of neovascular area in corneal whole mounts was performed using computer-assisted image analysis (Image-Pro Plus).

For adenoviral injections in the ovary, athymic nude rats were anaesthetized with isoflurane, and a 2–2.5-cm incision was made in the left dorsal area. The ovary was lifted and secured using non-traumatic forceps. Doses of 10^8 or 5×10^8 plaque-forming units (PFU) in 5–10 μ l of saline were injected by a gas-tight Hamilton syringe fitted with a 31-gauge needle (Hamilton). All procedures were performed in a biosafety cabinet with biosafety level 2 practices. For skeletal muscle (left gastrocnemius) or subcutaneous ear injections, athymic nude mice or rats were anaesthetized with isoflurane, the area was cleaned, and doses of 5×10^8 PFU of each virus preparation in a 50- μ l volume was injected per site. Animals in all adenovirus studies were killed 1 week after adenovirus administration. At necropsy, tissues were dissected and frozen or fixed for histological analysis.

Received 31 May; accepted 25 July 2001.

- Folkman, J. Angiogenesis in cancer, vascular, rheumatoid and other disease. *Nature Med.* **1**, 27–31 (1995).
- Ferrara, N. & Allitalo, K. Clinical applications of angiogenic growth factors and their inhibitors. *Nature Med.* **5**, 1359–1364 (1999).
- Carmeliet, P. Mechanisms of angiogenesis and arteriogenesis. *Nature Med.* **6**, 389–395 (2000).
- Yancopoulos, G. D. *et al.* Vascular-specific growth factors and blood vessel formation. *Nature* **407**, 242–248 (2000).
- Stewart, P. A. & Wiley, M. J. Developing nervous tissue induces formation of blood–brain barrier characteristics in invading endothelial cells: a study using quail-chick transplantation chimeras. *Dev. Biol.* **84**, 183–192 (1981).
- Aird, W. C. *et al.* Vascular bed-specific expression of an endothelial cell gene is programmed by the tissue microenvironment. *J. Cell Biol.* **138**, 1117–1124 (1997).
- Dellian, M., Witwer, B. P., Salehi, H. A., Yuan, F. & Jain, R. K. Quantitation and physiological characterization of angiogenic vessels in mice: effect of basic fibroblast growth factor, vascular endothelial growth factor/vascular permeability factor, and host microenvironment. *Am. J. Pathol.* **149**, 59–71 (1996).
- Simionescu, N. & Simionescu, M. in *Cell and Tissue Biology* (ed. Weiss, L.) 355–398 (Urban & Schwarzenberg, Baltimore, 1988).
- Palade, G. E., Simionescu, M. & Simionescu, N. Structural aspects of the permeability of the microvascular endothelium. *Acta Physiol. Scand.* (Suppl.) **463**, 11–32 (1979).
- Breier, G., Albrecht, U., Storrer, S. & Risau, W. Expression of vascular endothelial growth factor during embryonic angiogenesis and endothelial cell differentiation. *Development* **114**, 521–532 (1992).
- Monacci, W. T., Merrill, M. J. & Oldfield, E. H. Expression of vascular permeability factor/vascular endothelial growth factor in normal rat tissues. *Am. J. Physiol.* **264**, C995–C1002 (1993).
- Leung, D. W., Cachianes, G., Kuang, W. J., Goeddel, D. V. & Ferrara, N. Vascular endothelial growth factor is a secreted angiogenic mitogen. *Science* **246**, 1306–1309 (1989).
- Joubert, F. J. & Strydom, D. J. Snake venom. The amino acid sequence of protein A from *Dendroaspis polylepis polylepis* (black mamba) venom. *H.-S. Z. Physiol. Chem.* **361**, 1787–1794 (1980).

- Boisbouvier, J. *et al.* A structural homologue of colipase in black mamba venom revealed by NMR floating disulphide bridge analysis. *J. Mol. Biol.* **283**, 205–219 (1998).
- Wechsberger, C. *et al.* The mammalian homologues of frog Bv8 are mainly expressed in spermatocytes. *FEBS Lett.* **462**, 177–181 (1999).
- Glinka, A. *et al.* Dickkopf-1 is a member of a new family of secreted proteins and functions in head induction. *Nature* **391**, 357–362 (1998).
- Aravind, L. & Koonin, E. V. A colipase fold in the carboxy-terminal domain of the Wnt antagonists—the Dickkops. *Curr. Biol.* **8**, 477–478 (1998).
- Klagsbrun, M. Mediators of angiogenesis: the biological significance of basic fibroblast growth factor (bFGF)-heparin and heparan sulfate interactions. *Semin. Cancer Biol.* **3**, 81–87 (1992).
- Zetter, B. R. Migration of capillary endothelial cells is stimulated by tumour-derived factors. *Nature* **285**, 41–43 (1980).
- Arbiser, J. L. *et al.* Oncogenic H-ras stimulates tumor angiogenesis by two distinct pathways. *Proc. Natl. Acad. Sci. USA* **94**, 861–866 (1997).
- Roberts, W. G. & Palade, G. E. Neovascularization induced by vascular endothelial growth factor is fenestrated. *Cancer Res.* **57**, 765–772 (1997).
- Esser, S. *et al.* Vascular endothelial growth factor induces endothelial fenestrations in vitro. *J. Cell Biol.* **140**, 947–959 (1998).
- Gospodarowicz, D., Lepine, J., Massoglia, S. & Wood, I. Comparison of the ability of basement membranes produced by corneal endothelial and mouse-derived endodermal PF-HR-9 cells to support the proliferation and differentiation of bovine kidney tubule epithelial cells in vitro. *J. Cell Biol.* **99**, 947–961 (1984).
- Semenza, G. L. HIF-1: mediator of physiological and pathophysiological response to hypoxia. *J. Appl. Physiol.* **88**, 1474–1480 (2000).
- Dor, Y., Porat, R. & Keshet, E. Vascular endothelial growth factor and vascular adjustments to perturbations in oxygen homeostasis. *Am. J. Physiol.* **280**, C1367–C1374 (2001).
- Collin, O. & Bergh, A. Leydig cells secrete factors which increase vascular permeability and endothelial cell proliferation. *Int. J. Androl.* **19**, 221–228 (1996).
- Basset, D. L. The changes in the vascular pattern of the ovary of the albino rat during the estrous cycle. *Am. J. Anat.* **73**, 251–278 (1943).
- Ravindranath, N., Little-Ihrig, L., Phillips, H. S., Ferrara, N. & Zeleznik, A. J. Vascular endothelial growth factor messenger ribonucleic acid expression in the primate ovary. *Endocrinology* **131**, 254–260 (1992).
- Goldzih, J. W. & Green, J. A. The polycystic ovary. I. Clinical and histologic features. *J. Clin. Endocrinol. Metab.* **22**, 325–332 (1962).
- Phillips, H. S., Hains, J., Leung, D. W. & Ferrara, N. Vascular endothelial growth factor is expressed in rat corpus luteum. *Endocrinology* **127**, 965–967 (1990).
- Pettersson, A. *et al.* Heterogeneity of the angiogenic response induced in different normal adult tissues by vascular permeability factor/vascular endothelial growth factor. *Lab. Invest.* **80**, 99–115 (2000).
- Ferrara, N. *et al.* Heterozygous embryonic lethality induced by targeted inactivation of the VEGF gene. *Nature* **380**, 439–442 (1996).
- Gerber, H. P. *et al.* VEGF is required for growth and survival in neonatal mice. *Development* **126**, 1149–1159 (1999).
- Ferrara, N. *et al.* Vascular endothelial growth factor is essential for corpus luteum angiogenesis. *Nature Med.* **4**, 336–340 (1998).
- Gerber, H. P. *et al.* VEGF couples hypertrophic cartilage remodeling, ossification and angiogenesis during endochondral bone formation. *Nature Med.* **5**, 623–628 (1999).
- Kim, K. J. *et al.* Inhibition of vascular endothelial growth factor-induced angiogenesis suppresses tumour growth in vivo. *Nature* **362**, 841–844 (1993).
- Aiello, L. P. *et al.* Vascular endothelial growth factor in ocular fluid of patients with diabetic retinopathy and other retinal disorders. *N. Engl. J. Med.* **331**, 1480–1487 (1994).
- Ferrara, N. VEGF: An update on biological and therapeutic aspects. *Curr. Opin. Biotech.* **11**, 617–624 (2000).
- Ryan, A. M. *et al.* Preclinical safety evaluation of rhuMabVEGF, an antiangiogenic humanized monoclonal antibody. *Toxicol. Pathol.* **27**, 78–86 (1999).
- Miao, H. Q. *et al.* Neuropilin-1 mediates collapsin-1/semaphorin III inhibition of endothelial cell motility: Functional competition of collapsin-1 and vascular endothelial growth factor-165. *J. Cell Biol.* **146**, 233–242 (1999).
- Schweitz, H., Pacaud, P., Diocot, M., Moinier, D. & Ladzunski, M. MIT(1), a black mamba intestinal toxin with a new and highly potent activity on intestinal contraction. *FEBS Lett.* **461**, 183–188 (1999).
- Phillips, H. S., Hains, J. M., Laramie, G. R., Rosenthal, A. & Winslow, J. W. Widespread expression of BDNF but not NT3 by target areas of basal forebrain cholinergic neurons. *Science* **250**, 290–294 (1990).
- Melton, D. A. *et al.* Efficient in vitro synthesis of biologically active RNA and RNA hybridization probes from plasmids containing a bacteriophage SP6 promoter. *Nucleic Acids Res.* **12**, 7035–7056 (1984).
- Gille, H. *et al.* Analysis of biological effects and signaling properties of Flt-1 (VEGFR-1) and KDR (VEGFR-2). A reassessment using novel receptor-specific VEGF mutants. *J. Biol. Chem.* **276**, 3222–3230 (2001).
- Krupnik, V. E. *et al.* Functional and structural diversity of the human Dickkopf gene family. *Gene* **238**, 301–313 (1999).
- van Tilbeurgh, H., Sarda, L., Verger, R. & Cambillau, C. Structure of the pancreatic lipase–procolipase complex. *Nature* **359**, 159–162 (1992).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank the oligonucleotide-synthesis, DNA-sequencing and protein-purification groups; D. Tran and G. Bennett for protein iodination; P. Schow for cell sorting; M. van Hoy for necropsy; and R. Taylor and P. Tobin for histology. We also thank H. Chen for initial observations and J. Singh and L. Yu for cell culture and transfections.

Correspondence and requests for materials should be addressed to N.F. (e-mail: nf@gene.com).

Characterization of extrasolar terrestrial planets from diurnal photometric variability

E. B. Ford*, S. Seager† & E. L. Turner*

* Princeton University Observatory, Peyton Hall; and † Institute for Advanced Study, Einstein Drive, Princeton, New Jersey 08540, USA

The detection of massive planets orbiting nearby stars has become almost routine^{1,2}, but current techniques are as yet unable to detect terrestrial planets with masses comparable to the Earth's. Future space-based observatories to detect Earth-like planets are being planned. Terrestrial planets orbiting in the habitable zones of stars—where planetary surface conditions are compatible with the presence of liquid water—are of enormous interest because they might have global environments similar to Earth's and even harbour life. The light scattered by such a planet will vary in intensity and colour as the planet rotates; the resulting light curve will contain information about the planet's surface and atmospheric properties. Here we report a model that predicts features that should be discernible in the light curve obtained by low-precision photometry. For extrasolar planets similar to Earth, we expect daily flux variations of up to hundreds of per cent, depending sensitively on ice and cloud cover as well as seasonal variations. This suggests that the meteorological variability, composition of the surface (for example, ocean versus land fraction) and rotation period of an Earth-like planet could be derived from photometric observations. Even signatures of Earth-like plant life could be constrained or possibly, with further study, even uniquely determined.

NASA and ESA are now considering two ambitious space missions—TPF³ and Darwin⁴ respectively—to detect and characterize terrestrial planets orbiting nearby Sun-like stars. Although very different designs are being considered^{3,4,5,6}, all have the goal of spectroscopic characterization of the atmospheric composition, and in particular the detection of gases that are important for or caused by life on Earth, including O₂, O₃, CO₂, CH₄ and H₂O (ref. 7). A mission capable of measuring these spectral features would necessarily obtain sufficient signal-to-noise ratios to measure photometric variability. Many other important properties of an extrasolar planet could be derived from photometric measurements; this is the motivation for the investigation presented here. Photometric variability would be especially valuable for studying many planets quickly or a planet that is too dim for spectroscopic studies. Moreover, the photometric variability could be monitored concurrently with a spectroscopic investigation, as was done for the transiting extrasolar giant planet of HD209458 (ref. 8).

We have developed a code which calculates the total light scattered by an extrasolar planet towards an observer. The code performs Monte Carlo integrations, with single scattering, over a spherical planet using a map that specifies the scattering surface type at each point on the sphere; the code also uses a set of wavelength-dependent bidirectional reflectance distribution functions that specify the probability that light incident from one direction will scatter into another direction for each type of scattering surface (refs 9–19; see also B. Rock and J. Salisbury, <http://speclib.jpl.nasa.gov/archive/jhu/becknic/vegetation/txt/deciduous.txt>). The observed flux also depends on the viewing geometry which is specified by the phase angle (the angle between the star, planet and observer), obliquity (the direction of the planet's rotation axis), orbital inclination relative to the line of sight, and time of day (t). Although the geography and climate of an Earth-like extrasolar planet are completely unknown, we can attempt to calculate what an optical or

near-infrared light curve of Earth would look like if viewed from a very large distance. Simple variations about this basic reference model then allow us to explore reasonable possibilities for other Earth-like planets. We use a map of Earth from a one square degree satellite surface map which classifies each pixel as permanent ice, dirty/temporary ice, ocean, forest, brush, or desert²⁰. We consider cloudy models separately, using the scattering properties of Earth clouds²¹. We focus our attention on quadrature (a phase angle of 90°) for which the planet–star separation is largest and the observational constraints thus least severe.

The diurnal light curve from our cloud-free Earth model has variation as high as 150% (Fig. 1). The significant intensity variation is due to the facts that surfaces have different albedos and a relatively small part of the visible hemisphere dominates the total flux from an unresolved planet. For example, at quadrature less than 10% of the surface often produces more than 50% of the total reflected light. In our cloud-free Earth model the variability is primarily due to land (lower specular reflection) and ocean (higher) rotating in and out of view. The peak in the light curve at $t = 0.5$ day is caused by the high albedo of sand from the Sahara desert, and the dip at $t = 0.8$ day is caused by South America rotating into the location where there is usually specular reflection off the ocean. Ice, sand, oceans and vegetation can all produce significant features in the rotational light curve, but distinguish themselves by their colours (for example, ice is very nearly grey, whereas sand has an albedo which rises by about 30% between 450 nm and 750 nm).

On Earth clouds are extremely important to the reflected light curve both because of their very high albedos and their short timescales of formation, motion and dissipation. Figure 2 shows the diurnal light curve of our Earth model with daily, seasonal and annual average cloud coverage maps from the ISCCP database (ref. 22 and see <http://isccp.giss.nasa.gov/>). Here the intensity

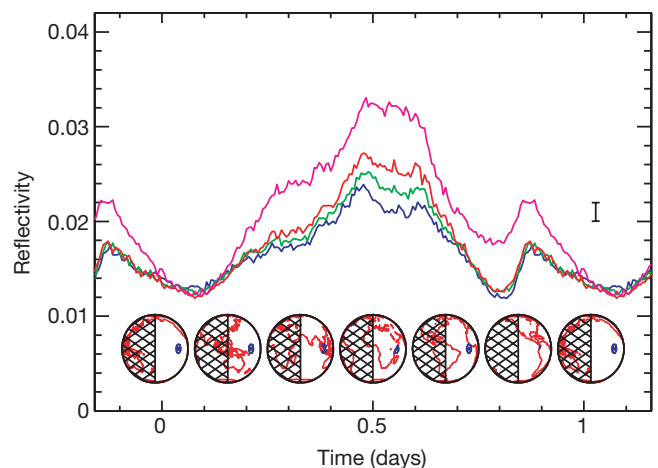


Figure 1 Rotational light curve for a cloud-free Earth model. The light curve is sampled at four-minute intervals and the Poisson noise is due to Monte Carlo statistics in our calculations. The pink, red, green and blue curves correspond to wavelengths of 750, 650, 550 and 450 nm, and their differences reflect the wavelength-dependent albedo of different surface components. The difference in magnitude of the star and planet is $\Delta m = 22.60 + 5 \log(r) - 5 \log(R_p) - 2.5 \log(R)$, where r is the star–planet separation in astronomical units, R_p is the radius of the planet in Earth radii, and R is the reflectivity normalized to a Lambert disk at a phase angle of 0° which is plotted on the y-axis. We note that a different phase angle affects R when a larger or smaller fraction of the disk is illuminated. The images below the light curve show the viewing geometry (the cross-hatched region is not illuminated), a map of the Earth (red), and the region of specular reflection from the ocean (blue). At $t = 0.5$ day, the Sahara desert is in view, causing a large peak in the light curve owing to the reflectivity of sand, which is especially high in the near-infrared (pink curve). The error bars in all figures show estimates of the photometric accuracy of a terrestrial planet finder mission.

variation is due to the contrast between clouds and land or ocean. Clouds tend to raise the overall brightness and variability, reducing the fractional variation in the reflected light curve compared to the cloud-free case (Fig. 1). On Earth cloud patterns cause a variation of about 20% within a typical day. Cloud patterns can be coherent over several days (Fig. 2a), allowing the rotational period of the Earth to be measured from its light curve and thus the light curve to be averaged over many days. Light curves of Earth with seasonal and annual cloud coverage maps (Fig. 2b) demonstrate that average cloud patterns for Earth vary with seasons. These seasonal light curves reflect the fact that there are locations that are almost always cloudy (such as the Amazon Basin) and regions that are virtually cloud free (such as the Sahara desert) at different times of the year.

We could learn about Earth's unresolved rotationally modulated flux from a satellite from Earth or by observing the reflection of the Earth from the dark side of the moon (Earthshine). Goode *et al.*²³ have measured Earthshine for about 200 days since 1998. Our model agrees with their measurements for both the mean reflectivity (0.086 for the model versus 0.092 that we observed) and fractional variance (13% model versus 15% observed) at a phase angle of between 80° and 100°. Their observations are made from a single location, limiting their range of viewing geometries. Monitoring Earthshine from several different longitude locations will allow for more complete daily coverage.

While the Earthshine observations²³ affirm the accuracy of our Earth model, they cannot address the light curves of extrasolar planets which will not be identical to those of Earth. The rotational period, the continental fraction and arrangement could be completely different. In addition, the cloud fraction and spatial and temporal distribution depends on many variables such as the location of continents and obliquity. Predicting cloud coverage is very difficult for global climate models, so we use cloud-free models to consider plausible Earth-like planets by varying the surface map

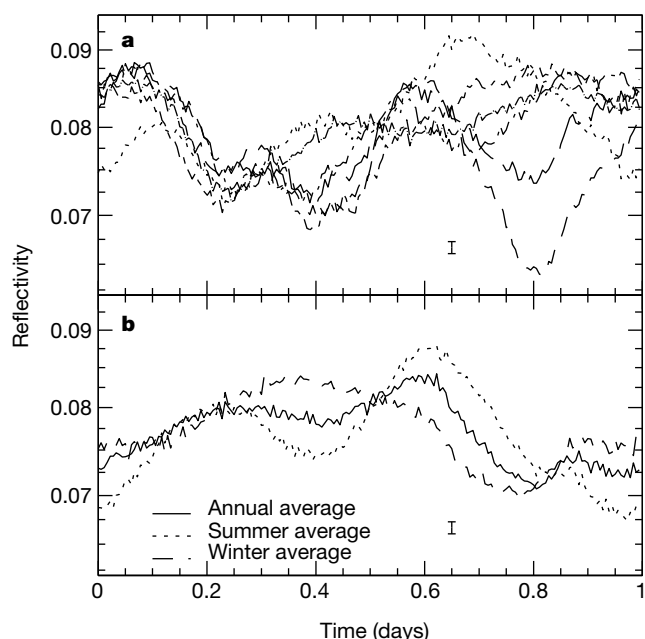


Figure 2 Rotational light curves for model Earth with clouds. Model daily light curves are shown at 550 nm for our Earth model with clouds, as viewed from quadrature. **a**, Theoretical light curves using cloud cover data from satellite measurements taken on six consecutive days (13–18 April 1986). **b**, Theoretical light curves for Earth using seasonal (dotted and dashed lines) and annual (solid line) average cloud cover (averaged over 8 years). Using actual cloud data allows us to accurately model the Earth, but is not applicable to extrasolar terrestrial planets.

(Fig. 3). Figure 3 shows daily light curves for planets with several different surface maps which illustrate that models with qualitatively different surfaces produce distinctly different light curves. For example, an Earth-like planet whose land was covered by ice (Fig. 3a) or thick forests (Fig. 3b) would have much larger amplitude variations than the Earth's. Varying the fractional ocean coverage (Fig. 3c) affects both the normalization and the variability of the light curve. Whereas Fig. 3c and d demonstrates that an extrasolar planet's diurnal light curve may contain information about the planet's surface, the full inverse problem of obtaining a unique determination of surface features from a light curve may be intractable and will certainly require much more investigation. For example, changing the obliquity of our cloudless Earth model also causes significant changes in the light curve (Fig. 3d). Nevertheless, a planet's light curve will clearly place constraints on its surface properties and climate in that many possible models would be ruled out by any specific set of photometric data.

In addition to studying 'geological' surface features we might learn about biological features encoded on the planet's surface. On Earth vegetation has a dramatic sudden rise in albedo by almost an order of magnitude at around 750 nm, known as the 'red edge'. Vegetation has evolved this strong reflection as a cooling mechanism to prevent overheating which would cause chlorophyll to degrade. Although the albedo of sand also increases towards the infrared, the red-edge vegetation signature is more rapid and may be detectable from the unresolved Earth. (See also the Galileo observations of part of Earth²⁴.) We cannot necessarily expect to find Earth-like

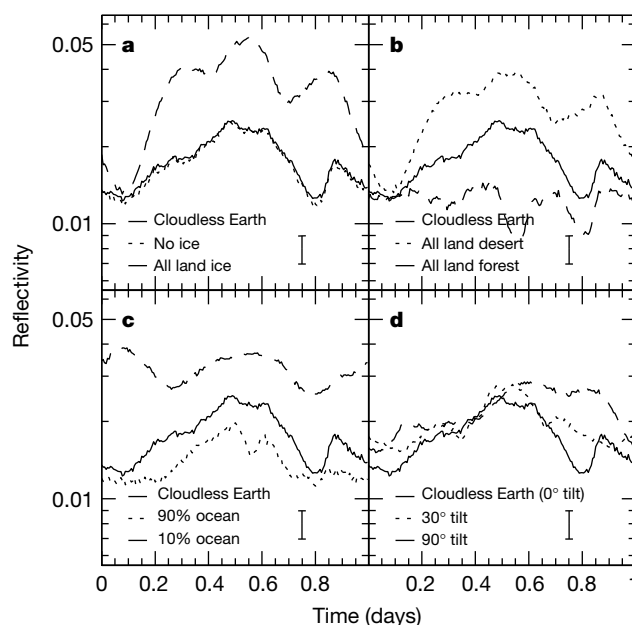


Figure 3 Rotational light curves for cloudless Earth-like models. Model daily light curves are shown at 550 nm for an Earth-like planet, as viewed from quadrature. The light curves shown in **a–c** are generated using variations on our cloud-free Earth model, which is presented as the solid black line in each panel for reference (corresponding to the green curve in Fig. 1). These models do not include clouds because the cloud coverage would differ for each Earth-variant in a way not yet known. **a**, Light curves for Earth-like planets for which all the land is covered with ice (dashed line) and none of the land is covered with ice (dotted line). **b**, Light curves for Earth-like planets for which all the land is covered with thick forests (dashed line) and all of the land is covered with desert (dotted line). **c**, Light curves for Earth-like planets for which the land and oceans have been recursively expanded and contracted along their perimeters to achieve the desired fractional ocean coverage. **d**, Light curves for our cloudless Earth model for different obliquities. The north pole has been tilted towards the Sun by 0° (solid line), 30° (dotted line), and 90° (dashed line).

vegetation on extrasolar terrestrial planets, but photometric measurements in different colours may be able to detect a unique signature, different from any known surface features or atmospheric constituents on Earth or other solar system planets.

We expect the diurnal rotational variation of an Earth-like planet to be lower in the mid-infrared (flux variation of a few per cent) than in the optical because the surface temperature does not vary as much as surface albedo across the Earth. For a planet with non-zero obliquity the mid-infrared seasonal flux variation should be larger than the mid-infrared rotational variation because of seasonal temperature variation. In addition, optical multicolour photometry and polarization as a function of phase angle (on an orbital timescale) could help constrain the type of atmospheric scatterers and the particle size distribution or detect a large smooth specular surface such as an ice- or an ocean-covered planet. Alternatively, the variation on an orbital timescale may be overwhelmed by variations indicating strong seasonal changes in the atmosphere.

Theoretical light curves of the unresolved Earth have a variation of 10–20%. A terrestrial planet finder that can measure 5% optical variation could detect weather, the rotational period, and seasonal changes in the cloud pattern of a planet like our Earth. If the surface were to contribute most of the scattered light rather than the clouds, then planets with different surface features would show very different diurnal light curves. This is in dramatic contrast to planets like Venus which would show almost no diurnal variability. Thus, we expect the diurnal light curve of an extrasolar Earth-like planet to contain detectable features encoding information about its physical and perhaps even biological properties. □

Received 17 April; accepted 9 July 2001.

1. Marcy, G. W., Cochran, W. D. & Mayor, M. in *Protostars and Planets IV* (eds Mannings, V., Boss, A. P., Russell, S. S.) 1285–1311 (Univ. Arizona Press, Tucson, 2000).
2. Butler, R. P., Marcy, G. W., Williams, E., McCarthy, C. & Dopsanjh, P. Attaining Doppler precision of 3 m s^{-1} . *Proc. Astron. Soc. Pacif.* **108**, 500–509 (1996).
3. Beichman, C. A., Woolf, N. J. & Lindensmith, C. A. The terrestrial planet finder (TPF). *NASA/JPL Publ.* **99-3**, 1–22 (1999).
4. Darwin the infrared space interferometer—concept and feasibility study report. Technical Report ESA-SCI(2000)12, 1–85 (European Space Agency, 2000).
5. Nisenson, P. & Papaliolios, C. Detection of earth-like planets using apodized telescopes. *Astrophys. J.* **548**, L201–205 (2001).
6. Spergel, D. N. A new pupil for detecting extrasolar planets. *Appl. Optics* (submitted); also preprint astro-ph/0101142 at (xxx.lanl.gov) (2001).
7. Schindler, T. L. & Kasting, J. F. Synthetic spectra of simulated terrestrial atmospheres containing possible biomarker gases. *Icarus* **145**, 262–271 (2000).
8. Brown, T. M., Charbonneau, D., Gilliland, R. L., Noyes, R. W. & Burrows, A. Hubble Space Telescope time-series photometry of the transiting planet of HD 209458. *Astrophys. J.* **552**, 699–709 (2001).
9. Hapke, B. *Theory of Reflectance and Emittance Spectroscopy* (Cambridge Univ. Press, Cambridge, 1993).
10. Bukata, R. P., Jerome, J. H., Kondratyev, K. Ya. & Pozdnyakov, D. V. *Optical Properties and Remote Sensing of Inland and Coastal Waters* 37–102 (CRC Press, New York, 1995).
11. Vokrouhlicky, D. & Farinella, P. Specular reflection of sunlight from wavy ocean surfaces and the albedo effect on satellite orbits I. A statistical model. *Astron. Astrophys.* **293**, 307–322 (1995).
12. Jerome, J. H., Bukata, R. P. & Bruton, J. E. Utilizing the components of vector irradiance to estimate the scalar irradiance in natural waters. *Appl. Optics* **27**, 4012–4018 (1988).
13. Kirk, J. T. O. Volume scattering function, average cosines, and the underwater light field. *Limnol. Oceanogr.* **36**, 455–467 (1991).
14. Bricaud, A., Babin, M., Morel, A. & Claustre, H. Variability in the chlorophyll-specific absorption coefficients of natural phytoplankton: analysis and parameterization. *J. Geophys. Res.* **100**, 13321–13332 (1995).
15. Kondratyev, K. Ya. *Radiation in the Atmosphere* (Academic, New York, 1969).
16. Verbeke, A. J. & Veverka, J. Scattering properties of natural snow and frost: comparison with icy satellite photometry. *Icarus* **88**, 418–428 (1990).
17. Myrneni, R. B. et al. Optical remote sensing of vegetation: modeling, caveats, and algorithms. *Remote Sens. Environ.* **51**, 169–188 (1995).
18. Jacquemoud, S., Baret, F. & Hanocq, J. F. Modeling spectral and bidirectional soil reflectance. *Remote Sens. Environ.* **41**, 123–132 (1992).
19. Woessner, P. & Hapke, B. Polarization of light scattered by clover. *Remote Sens. Environ.* **21**, 243–261 (1987).
20. DeFries, R. S. & Townshend, J. R. G. NDVI-derived land cover classification at a global scale. *Int. J. Remote Sens.* **15**, 3567–3586 (1994).
21. Hovenier, J. W. & Hage, J. I. Relations involving the spherical albedo and other photometric quantities of planets with thick atmospheres. *Astron. Astrophys.* **214**, 391–401 (1989).
22. Rossow, W. B. & Schiffer, R. A. ISCCP cloud data products. *Bull. Am. Meteorol. Soc.* **72**, 2–20 (1991).
23. Goode, P. R. et al. Earthshine Observations of the Earth's Reflectance. *J. Geophys. Res. Lett.* **28**, 1671–1674 (2001).

24. Sagan, C., Thompson, W. R., Carlson, R., Gurnett, D. & Hord, C. A. Search for life on earth from the Galileo spacecraft. *Nature* **365**, 715 (1993).

Acknowledgements

We thank the members of the Ball Aerospace Terrestrial Planet Finder team, B. Soden, A. Broccoli and G. Williams for discussions, and B. Soden for assistance with the ISCCP database. We wish to thank W. B. Rossow, the Goddard Institute for Space Studies, the Goddard Space Flight Center, and NASA for the production and distribution of this data set. E.B.F. would like to acknowledge support under a National Science Foundation Graduate Fellowship and S.S. is supported by the W. M. Keck Foundation.

Correspondence and requests for materials should be addressed to E.B.F. (e-mail: eford@astro.princeton.edu).

Stimulated emission of polarization-entangled photons

A. Lamas-Linares, J. C. Howell & D. Bouwmeester

Centre for Quantum Computation, Clarendon Laboratory, University of Oxford, Parks Road, Oxford OX1 3PU, UK

Entangled photon pairs—discrete light quanta that exhibit non-classical correlations—play a crucial role in quantum information science (for example, in demonstrations of quantum non-locality^{1–7}, quantum teleportation^{8,9} and quantum cryptography^{10–12,31}). At the macroscopic optical-field level non-classical correlations can also be important, as in the case of squeezed light¹³, entangled light beams^{14,15} and teleportation of continuous quantum variables¹⁶. Here we use stimulated parametric down-conversion to study entangled states of light that bridge the gap between discrete and macroscopic optical quantum correlations. We demonstrate experimentally the onset of laser-like action for entangled photons, through the creation and amplification of the spin-1/2 and spin-1 singlet states consisting of two and four photons, respectively. This entanglement structure holds great promise in quantum information science where there is a strong demand for entangled states of increasing complexity.

As the acronym LASER (light amplification by stimulated emission of radiation) indicates, polarization-entangled laser operation would mean that a (spontaneously created) photon pair in two polarization-entangled modes stimulate, inside a nonlinear gain medium, the emission of additional pairs. As a gain medium we consider type-II parametric down-conversion¹⁷. A simplified interaction hamiltonian^{18,19} for the nonlinear interaction between a classical pump field and two polarization-entangled modes a and b is given by

$$\hat{H}_{\text{int}} = e^{i\phi} \kappa \hat{K}^{\dagger} + e^{-i\phi} \kappa \hat{K} \quad (1)$$

where $\hat{K}^{\dagger} \equiv (\hat{a}_H^{\dagger} \hat{b}_V^{\dagger} - \hat{a}_V^{\dagger} \hat{b}_H^{\dagger})$ and $\hat{K} \equiv (\hat{a}_H \hat{b}_V - \hat{a}_V \hat{b}_H)$ are the creation and annihilation operators of polarization-entangled photon pairs in modes a and b . Horizontal and vertical polarization are represented by H and V, and κ is a real-valued coupling coefficient. When acting on the vacuum state the time evolution operator $\hat{U} = \exp(i\hat{H}t/\hbar)$ yields:

$$|\psi\rangle \propto \sum_{n=0}^{\infty} (\tanh \tau)^n \sum_{m=0}^n (-1)^m |n-m, m; m; n-m\rangle \quad (2)$$

where $\tau \equiv \kappa t/\hbar$ is the interaction parameter. The first and second positions in the ket indicate respectively the number of horizontal ($n-m$) and vertical (n) photons in mode a , and the third and fourth positions indicate the corresponding numbers for mode b . This state represents the general output of type-II parametric down-

conversion, but for all experiments reported to date, τ is so small that mainly the first-order term ($n = 1$) has been taken into account and only a few experiments and proposals have addressed second-order terms^{7,20–24}. By analogy with a conventional laser, the idea of an entangled-photon laser is to increase τ using a resonator around the gain medium, which enhances the emission of the higher-order terms in equation (2). The state shown in equation (2) has the following features.

First, modes a and b are entangled in photon number because, for any n , the number of photons in each mode is identical. The photon (pair) number distribution is shown in Fig. 1 for increasing average photon (pair) number output per pulse $\langle n \rangle$. The shifting of the maximum and the broadening of the distribution for higher values of $\langle n \rangle$ resembles the coherent state photon-number distribution as produced by conventional lasers. These features are explained by the fact that stimulated emission—originating from the boson statistics of photons—favors amplification of higher over lower photon-number terms.

The second important property of the state in equation (2) is that the set of terms for each n form a maximally entangled state in polarization. The normalized 1-pair term is the rotationally symmetric Bell state (singlet spin-1/2):

$$|\psi\rangle = \frac{1}{\sqrt{2}}(|1, 0; 0, 1\rangle - |0, 1; 1, 0\rangle) \quad (3)$$

The normalized 2-pair term is given by

$$|\psi\rangle = \frac{1}{\sqrt{3}}(|2, 0; 0, 2\rangle - |1, 1; 1, 1\rangle + |0, 2; 2, 0\rangle) \quad (4)$$

and represents the singlet spin-1 state. Similar to the spin-1/2 case, the rotational symmetry arises from the relative phase relations and the equal weights of the terms. In general, the n -pair term has the properties of a singlet spin- $n/2$ state. The rotational symmetry of the full state can be shown by expressing the n -pair terms in any other basis, and verifying that the same expressions are obtained. The crucial role of stimulated emission is to provide for each n equally weighted terms. In principle, a photon counting measurement on state (2) (either in mode a or b) performs a projection onto a certain singlet spin- $n/2$ state. Subsequently, this maximally entangled state can be explored for quantum information tasks. In practice, in quantum optics experiments where the fragile photons are in general destroyed by any measurement, the projection and the exploration of the state are performed simultaneously. This procedure, usually referred to as post-selection, has proved to be very useful: for example, for demonstrations of quantum teleportation⁸,

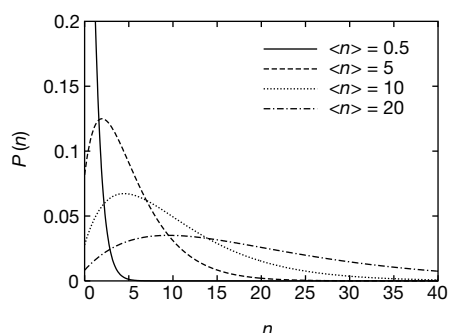


Figure 1 The photon number (pair) distribution, $P(n)$, arising from stimulated parametric down-conversion shifts its peak and broadens as the mean number of photons increases. This indicates that for increasing interaction strength (gain), terms with higher numbers of photons obtain a larger amplification factor compared to lower terms, which is a feature of laser operation.

quantum cryptography^{10–12,31}, and three-particle Greenberger–Horne–Zeilinger (GHZ) correlations^{7,21}, and for novel optical quantum computation schemes²⁵. Here we use post-selection to demonstrate stimulated entanglement by measuring 2- and 4-photon properties of state (2) for increasing values of τ .

The set-up used to demonstrate stimulated entanglement is illustrated in Fig. 2. A 120-fs pump pulse at 390 nm wavelength (with a repetition rate of 80 MHz) passes through a β -barium borate (BBO) crystal and creates pairs of polarization-entangled photons in spatially distinct modes a and b . The experimental parameters are chosen such that (to first order) the singlet photon-pair state shown in equation (3) is created. Initially modes a and b are in the vacuum state and the photon pairs are spontaneously created. The fact that modes a and b geometrically diverge, and that horizontally and vertically polarized photons experience different crystal parameters, limits the useful crystal length²⁶ and thereby prohibits an efficient stimulated emission process. To obtain significant stimulated emission we redirect the spontaneously created photon pairs into the crystal at the same time (tuned by a delay on mirror M3) as the reflected pump pulse passes through the crystal a second time. Provided that the feedback loop for the photon pairs is polarization independent, which is obtained by using a bow-tie folded geometry including a $\lambda/2$ waveplate that exchanges H and V polarizations, optimum conditions for stimulated emission of photon pairs can be established. As stimulated emission can be seen as a constructive multi-particle interference effect, and because the process of parametric down-conversion is sensitive to the phase of the pump, we should expect to observe an oscillation between stimulation and suppression of emission as function of the pump-pulse delay. The period of this oscillation corresponds to the optical frequency of the pump laser. In the region where the difference between the pump delay and the feedback loop is larger than the coherence length of the observed photons (determined by the 5-nm narrow-bandwidth filters in front of the single-photon detectors), no such interference pattern is expected.

To study the 2- and 4-photon entangled states we measure each term in equations (3) and (4) individually in two non-orthogonal polarization bases. The $|1, 0; 0, 1\rangle$ and the $|0, 1; 1, 0\rangle$ terms are measured in the desired bases by using a polarizer in front of a single-photon detector in each of the spatial modes a and b . The $|1, 1; 1, 1\rangle$ term is detected by the introduction of polarizing beam splitters in the appropriate basis in each mode followed by four single-photon detectors (Fig. 2). As we do not use multi-photon detectors, we can

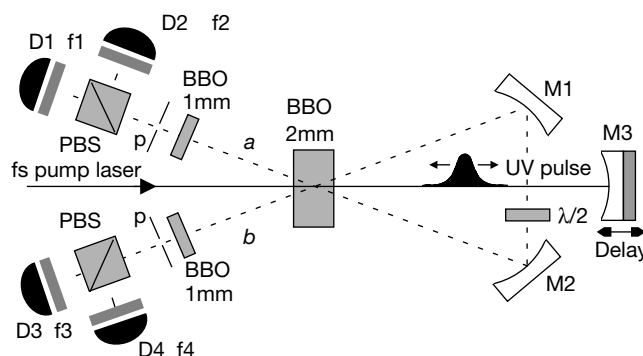


Figure 2 Experimental set-up. A frequency-doubled, mode-locked Ti:sapphire laser (80 MHz repetition rate, $\lambda = 390$ nm) pumps a 2-mm (BBO) β -barium borate crystal. Pinholes (p) perform spatial selection of the entangled modes. The pump is reflected onto itself by mirror M3 that is mounted on a computer-controlled translation stage. Mirrors M1 and M2 form the feedback loop, including a polarization rotation element ($\lambda/2$), for the entangled photons. Photon detection of the $|1, 1; 1, 1\rangle$ term in the H/V basis occurs at avalanche photodiodes D1–D4, after going through polarizing beam splitters (PBS) and 5-nm-bandwidth filters f1–f4. The role of the two extra 1-mm BBO crystals in modes a and b is to compensate for undesirable birefringent properties of the main crystal¹⁷.

only measure the $|2,0;0,2\rangle$ and the $|0,2;2,0\rangle$ term with a 0.25 probability using a combination of a polarizer, a 50-50 beam splitter and two single-photon detectors in each mode.

Quantitative predictions for the amplification of the individual terms in equations (3) and (4) resulting from the double-pass configuration are obtained by expanding the unitary evolution of the created light fields in the polarization-entangled photon-pair creation operator $\hat{K}^\dagger$. To second order in $\hat{K}^\dagger$ we obtain

$$\begin{aligned} \hat{U} = \hat{U}_2 \hat{U}_1 = 1 + e^{i\theta} \tau \hat{K}_2^\dagger + \tau \hat{K}_1^\dagger \\ + \frac{1}{2} e^{2i\theta} \tau^2 (\hat{K}_2^\dagger)^2 + \frac{1}{2} \tau^2 (\hat{K}_1^\dagger)^2 + e^{i\theta} \tau^2 \hat{K}_2^\dagger \hat{K}_1^\dagger \end{aligned} \quad (5)$$

where subscripts 1 and 2 refer to the first and second pass through

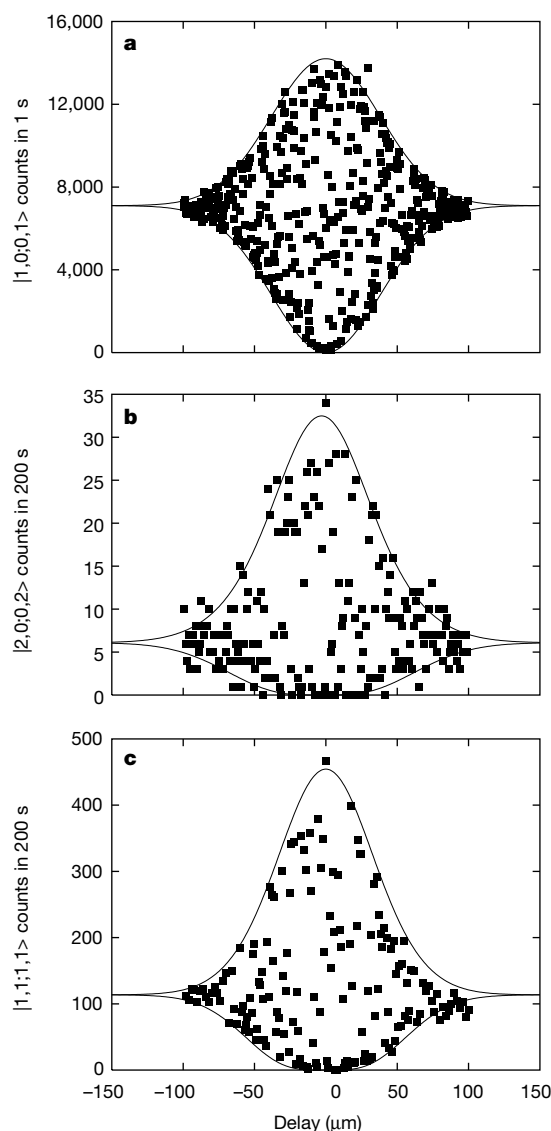


Figure 3 Experimental demonstration of stimulated entanglement. **a**, The 2-fold coincidence rate corresponding to the detection of the $|1,0;0,1\rangle$ term in the 45° rotated basis as function of the delay between the reflected pump and the entangled photons generated in the first pass through the crystal. Solid lines, theoretical fits to the envelope of the curve as the degree of overlap varies. **b**, The 4-fold coincidence rate corresponding to the detection of the $|2,0;0,2\rangle$ term in the 45° rotated basis, and **c**, the $|1,1;1,1\rangle$ in the H/V basis. The effect of stimulated emission is apparent in the increase of the number of 4-fold coincidences at zero delay of a factor of 5.3 and 4.0 for the $|2,0;0,2\rangle$ and $|1,1;1,1\rangle$ terms, respectively (see text). The difference in rates between the two 4-photon graphs is due to the probabilistic detection and extra elements introduced to measure the $|2,0;0,2\rangle$ term.

the crystal. The relative phase θ between the first and second pass of the pump pulse is tunable by the translation of mirror M3 (Fig. 2). We discuss two limits: when $\hat{K}_1^\dagger = \hat{K}_2^\dagger$, and when $\hat{K}_1^\dagger$ is distinguishable from $\hat{K}_2^\dagger$. The first case applies at zero delay where efficient phase-sensitive stimulated emission occurs. From equation (5) it follows that doubling the value of the interaction parameter τ results in an increase in probability for the 2-photon terms from τ^2 to $4\tau^2$, and in an increase for the 4-photon terms from τ^4 to $16\tau^4$. We note that the 4-photon state has a four times larger amplification than the 2-photon states, which is characteristic of stimulated emission. The second case applies if the reflected pump-pulse delay between the two passes is not equal to the delay of the entangled photons in the feedback loop. In this case, there are simply two independent contributions to the 2-photon detection events, but there are several distinct contributions to the 4-photon detection events. Each single pass has a small probability (of τ^4) to create state (4). In addition, as current single-photon detectors do not have a high enough time resolution to distinguish between photons arriving from the first or second pass, there are spurious 4-fold coincidences from a combination of 2-photon states created in both passes. The spurious contributions to the $|2,0;0,2\rangle$ and the $|0,2;2,0\rangle$ detections will be τ^4 , and for the $|1,1;1,1\rangle$ detection will be $2\tau^4$.

We scan from the region where $\hat{K}_1^\dagger$ is completely distinguishable from $\hat{K}_2^\dagger$ into the region where $\hat{K}_1^\dagger = \hat{K}_2^\dagger$, while observing the intensity of the 2- and 4-photon terms. From the considerations above we expect—in the case of ideal stimulated emission—the terms in equation (3) to show a 2-fold increase, the middle term in equation (4) to show a 4-fold increase, and the other two terms in equation (4) to increase by a factor of $16/3 = 5.33$. Owing to the rotational symmetry of states (3) and (4), these predictions are basis independent.

Figure 3 shows our experimental data for the detection of the $|1,0;0,1\rangle$ (panel **a**) and the $|2,0;0,2\rangle$ (b) terms measured in the 45°

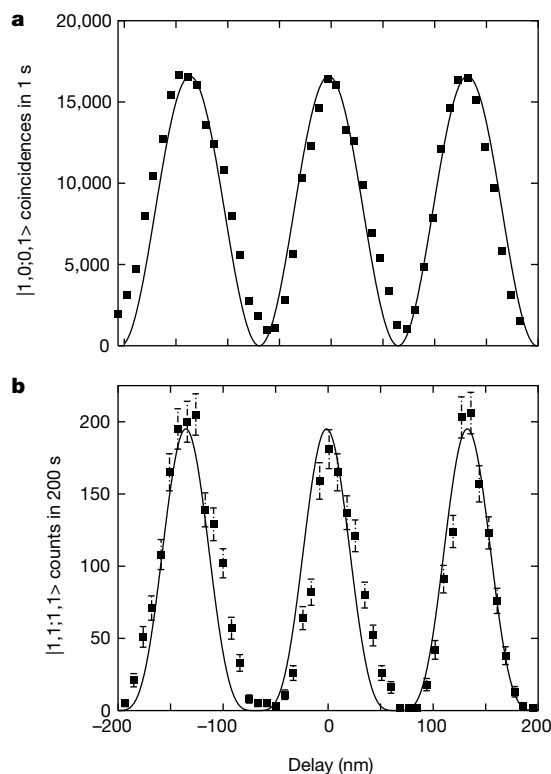


Figure 4 Two- and four-photon interference due to stimulated emission. A fine scan of the $|1,0;0,1\rangle$ (**a**) and $|1,1;1,1\rangle$ (**b**) terms in the zero-delay region shows optimum stimulation and suppression of the 2- and 4-fold coincidence probability.

rotated basis, and the $|1,1;1,1\rangle$ (c) term in the H/V basis. The solid curves are the envelopes of the oscillating functions giving the maximum and minimum theoretical values for the coincidence rates. The experimental data show an increase of 1.95 ± 0.10 for $|1,0;0,1\rangle$, 5.3 ± 0.6 for $|2,0;0,2\rangle$ and of 4.1 ± 0.3 for $|1,1;1,1\rangle$. These results are in good agreement with the predictions discussed above. Similar results have been obtained in the other bases and for the $|0,1;1,0\rangle$ and $|0,2;2,0\rangle$ terms, demonstrating the rotational invariance—that is, the spin-1/2 and spin-1 singlet structure—of states (3) and (4). Additional data indicate an amplification due to the second pass of 3.95 ± 0.10 for the 2-fold coincidences, and of 17 ± 2 for the 4-fold coincidences. This demonstrates the shifting of the photon-number pair distribution towards terms with higher photon numbers, a characteristic of stimulated emission. A final proof of stimulated emission—seen as a constructive interference process—is the phase-dependent emission probability shown in Fig. 4. This is a fine scan around the region of zero delay for the $|1,0;0,1\rangle$ (panel a) and the $|1,1;1,1\rangle$ (b) terms in the 45° rotated basis and H/V basis, respectively. The solid lines are fits to the theoretical predictions, which vary as $(1 + \cos\theta)$ for the 2-photon case, and as $(1 + \cos\theta)^2$ for the 4-photon case. The visibility of these interference fringes is in all cases above 97%.

In summary, we have pointed out that the rich entanglement structure—equivalent to a superposition of spin- $n/2$ singlet states—obtained by stimulated emission of the familiar rotationally symmetric Bell state; we have also demonstrated that stimulated emission of polarization-entangled photons can be achieved experimentally. Both the characteristic shifting of the photon pair distribution towards higher photon numbers, and the rotational symmetry of the 2- and 4-photon contribution, have been observed. The good agreement between the experimental data and the theory shows that the stringent indistinguishability requirements to obtain entanglement in a stimulated process using external resonators have been met. Although related theory and experiments on interference-enhanced emission of photon pairs and on photon injection into nonlinear crystals have been reported^{22,23,27,28}, our results constitute (to our knowledge) the first experimental demonstration of the onset of laser-like operation for entangled photons.

Using multi-pass amplification pumped by higher-intensity pulses it should be possible to produce rotationally symmetric multi-photon entangled states with an average photon (pair) number of the order of 100. As exploration of such states is based on post-selection, the challenge of creating them should go in parallel with the challenge of constructing low-loss transmission lines and high-efficiency multi-photon detectors. Although there are encouraging developments in low-loss optical fibres²⁹ and highly efficient multi-photon detectors³⁰, we will always have to face the situation of losing photons in the process of creating, transporting and analysing the desired state of equation (2). As this state is one large, complex, entangled state, one might think that the loss (or a measurement) of a single photon would destroy all the interesting properties of the state. On the contrary, the complex entanglement has the remarkable feature that the loss or measurement of one (or more) particles does not eliminate all the entanglement between the remaining particles. To illustrate this point, we focus our attention on the 4-photon state of equation (4), and consider a measurement of a photon in mode a in the H/V basis, with the measurement result being H. The state of the remaining three particles will be $(1/\sqrt{2})|1,0;0,2\rangle - |0,1;1,1\rangle$, which still contains non-maximal entanglement between modes a and b . A generalization to the actual loss of several photons from the full state (2) is currently under study.

The increased sharpness of the peaks in the interference pattern for the singlet of spin-1 (see Fig. 4) is potentially of interest for applications in metrology. An interferometric measurement of relative distance, for example, would benefit from the more precise

location of the maxima obtained by looking at the 4-fold (or higher) coincidence rates. Furthermore, the increased amount of entangled terms made available by stimulated emission and post-selection offer new possibilities for higher-bit-rate quantum cryptography. We consider that entanglement robustness—together with the rotational symmetry of the state created by stimulated polarization entanglement—opens the way to many applications in quantum information, and provides a powerful tool for studying the almost unexplored area between the discrete and the macroscopic optical quantum correlation experiments. □

Received 2 May; accepted 3 July 2001.

1. Clauser, J. F. & Shimony, A. Bell's theorem: experimental tests and implications. *Rep. Prog. Phys.* **41**, 1881–1927 (1978).
2. Aspect, A., Dalibard, J. & Roger, G. Experimental test of Bell's inequalities using time-varying analyzers. *Phys. Rev. Lett.* **49**, 1804–1807 (1982).
3. Ou, Z. Y. & Mandel, L. Violation of Bell's inequality and classical probability in a two-photon correlation experiment. *Phys. Rev. Lett.* **61**, 50–53 (1988).
4. Shih, Y. H. & Alley, C. O. New type of Einstein-Podolski-Rosen experiment using pairs of quanta produced by optical parametric down-conversion. *Phys. Rev. Lett.* **61**, 2921–2924 (1988).
5. Tittel, W., Brendel, J., Zbinden, H. & Gisin, N. Violation of Bell inequalities by photons more than 10 km apart. *Phys. Rev. Lett.* **81**, 3563–3566 (1998).
6. Weihs, G., Jennewein, T., Simon, C., Weinfurter, H. & Zeilinger, A. Violation of Bell's inequality under strict locality conditions. *Phys. Rev. Lett.* **81**, 5039–5043 (1998).
7. Pan, J.-W. *et al.* Experimental test of non-locality in three-photon Greenberger-Horne-Zeilinger entanglement. *Nature* **403**, 515–519 (2000).
8. Boumeester, D. *et al.* Experimental quantum teleportation. *Nature* **390**, 575–579 (1997).
9. Boschi, D. *et al.* Experimental realization of teleporting an unknown pure quantum state via dual classical and Einstein-Podolski-Rosen channels. *Phys. Rev. Lett.* **80**, 1121–1125 (1998).
10. Ekert, A., Rarity, J. G., Tapster, P. R. & Palma, G. M. Practical quantum cryptography based on two-photon interferometry. *Phys. Rev. Lett.* **69**, 1293–1295 (1992).
11. Jennewein, T., Simons, C., Weihs, G., Weinfurter, H. & Zeilinger, A. Quantum cryptography with entangled photons. *Phys. Rev. Lett.* **84**, 4729–4732 (2000).
12. Tittel, W., Brendel, J., Zbinden, H. & Gisin, N. Quantum cryptography using entangled photons and energy-time Bell states. *Phys. Rev. Lett.* **84**, 4737–4736 (2000).
13. Slusher, R. E. *et al.* Observation of squeezed states generated by four-wave mixing in an optical cavity. *Phys. Rev. Lett.* **55**, 2409–2412 (1985).
14. Wu, L. A., Kimble, H. J., Hall, J. L. & Wu, H. Generating squeezed states by parametric down-conversion. *Phys. Rev. Lett.* **57**, 2520–2523 (1986).
15. Ou, Z. Y., Pereira, S. F., Kimble, H. J. & Peng, K. C. Realization of the Einstein-Podolski-Rosen paradox for continuous variables. *Phys. Rev. Lett.* **68**, 3663–3666 (1992).
16. Furusawa, A. *et al.* Unconditional quantum teleportation. *Science* **282**, 706–709 (1998).
17. Kwiat, P. G. *et al.* New high-intensity source of polarization-entangled photon pairs. *Phys. Rev. Lett.* **75**, 4337–4341 (1995).
18. Kok, P. & Braunstein, S. L. Postselected versus non-postselected quantum teleportation using parametric down-conversion. *Phys. Rev. A* **61**, 042304-1–042304-10 (2000).
19. Simon, C., Weihs, G. & Zeilinger, A. Optical quantum cloning via stimulated emission. *Phys. Rev. Lett.* **84**, 2993–2996 (2000).
20. Ou, Z. Y., Rhee, J.-K. & Wang, L. J. Observation of four-photon interference with a beam splitter by pulsed parametric down-conversion. *Phys. Rev. Lett.* **83**, 959–962 (1999).
21. Bouwmeester, D. *et al.* Observation of three-photon Greenberger-Horne-Zeilinger entanglement. *Phys. Rev. Lett.* **82**, 1345–1349 (1999).
22. De Martini, F., Mussi, V. & Bovino, F. Schroedinger cat states and optimum universal quantum cloning by entangled parametric amplification. *Opt. Commun.* **179**, 581–589 (2000).
23. DeMartini, F. & Di Giuseppe, G. Multiparticle quantum superposition and stimulated entanglement by parity selective amplification of entangled states. *Z. Naturforsch.* **56**, 61–66 (2001).
24. Weinfurter, H. & Zukowski, M. Four photon entanglement from down conversion. Preprint quant-ph/0103049 at (<http://xxx.lanl.gov>) (2001).
25. Knill, E., Laflamme, R. & Milburn, G. J. A scheme for efficient quantum computation with linear optics. *Nature* **409**, 46–52 (2001).
26. Atature, M. *et al.* Partial distinguishability in femtosecond optical spontaneous parametric down-conversion. *Phys. Rev. Lett.* **83**, 1323–1326 (1999).
27. Herzog, T. J., Rarity, J. G., Weinfurter, H. & Zeilinger, A. Frustrated two-photon creation via interference. *Phys. Rev. Lett.* **72**, 629–632 (1994).
28. Milonni, P. W., Fearn, H. & Zeilinger, A. Theory of two-photon down-conversion in the presence of mirrors. *Phys. Rev. A* **53**, 4556–4566 (1996).
29. Cregan, R. F. *et al.* Single-mode photonic band gap guidance of light in air. *Science* **285**, 1537–1539 (1999).
30. Kim, J., Takeuchi, S. & Yamamoto, Y. Multiphoton detector using visible light photon counter. *Appl. Phys. Lett.* **74**, 902–904 (1999).
31. Naik, D. S., Peterson, C. G., White, A. G., Berghlund, A. J. & Kwiat, P. G. Entangled state quantum cryptography: eavesdropping on the Ekert protocol. *Phys. Rev. Lett.* **84**, 4733–4736 (2000).

Acknowledgements

We thank C. Mikkelsen, P. Varisco, W. Irvine, A. Ekert and J. Rarity for suggestions and experimental support. This work was supported by the EPSRC, the UK Defence Evaluation and Research Agency, and the European QuComm project.

Correspondence and requests for materials should be addressed to A.L.-L. (e-mail: a.lamas@qubit.org).

Mechanism for the destruction of H_3^+ ions by electron impact

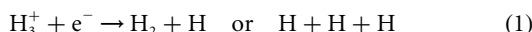
V. Kokoouline*, Chris H. Greene* & B. D. Esry†

* Department of Physics and JILA, University of Colorado, Boulder, Colorado 80309-0440, USA

† Department of Physics, Cardwell Hall, Kansas State University, Manhattan, Kansas 66506, USA

The rate at which the simplest triatomic ion (H_3^+) dissociates following recombination with a low-energy electron has been measured in numerous experiments^{1–10}. This process is particularly important for understanding observations of H_3^+ in diffuse interstellar clouds^{11–13}. But, despite extensive efforts^{14,15}, no theoretical treatment has yet proved capable of predicting the measured dissociative recombination rates at low energy, even to within an order of magnitude. Here we show that the Jahn–Teller symmetry-distortion effect^{16–19}—almost universally neglected in the theoretical description of electron–molecule collisions—generates recombination at a much faster rate than any other known mechanism. Our estimated rate constant overlaps the range of values spanned by experiments. We treat the low-energy collision process as a curve-crossing problem, which was previously thought inapplicable²⁰ to low-energy recombination in H_3^+ . Our calculation reproduces the measured propensity for three-body versus two-body breakup of the neutral fragments³, as well as the vibrational distribution⁴ of the H_2 product molecules.

Despite the key role of the H_3^+ ion in many astrophysical and chemical environments^{11–13}, the dissociative recombination (DR) of H_3^+ with a low-energy electron:



has remained poorly understood despite experimental progress since the early 1990s. The high DR rate measured in laboratory storage-ring experiments^{1–7} seems incompatible with the large amount of H_3^+ observed in diffuse interstellar clouds^{11–13}. Yet one model²¹ suggests that the high storage-ring rate can be understood in terms of the ‘clumpy’ nature of the intervening interstellar clouds responsible for the absorption of infrared radiation at the H_3^+ resonance lines. Another problem is that flowing-afterglow Langmuir probe (FALP) measurements obtain a DR rate that is one^{8,10} order of magnitude smaller (or even less⁹) than rates extracted from storage-ring experiments—and there is an orders-of-magnitude discrepancy between the storage-ring experiments^{1,5–7} and theoretical estimates^{14,15}. No satisfactory theoretical model has yet been able to predict a DR rate anywhere near as large as experiments measure for H_3^+ at sub-eV electron energies.

Here we develop a method that includes the non-perturbative Jahn–Teller coupling between the electronic and nuclear motions. Although a complete quantitative implementation of this method remains to be made, a preliminary implementation presented here allows us to understand at least the large value of the FALP experimental cross-section measurements^{8–10}. The theory also correctly predicts other details (within the approximations discussed below), such as the DR branching ratio³ ($\text{H} + \text{H} + \text{H})/(\text{H}_2 + \text{H})$, and the most probable vibrational quantum ($\nu \approx 5$) of the H_2 products⁴.

We treat the ionic vibrational dynamics on the H_3^+ electronic potential surface²² within the adiabatic hyperspherical approximation^{23–25}. The hyperspherical radius R is a useful collective coordinate for any few-body system: for three atoms of equal mass, the squared hyperradius is defined by $R^2 = \sqrt{3}(r_1^2 + r_2^2 + r_3^2)$, where r_1 , r_2 and r_3 are the distances of each nucleus from the centre

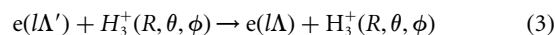
of mass. The adiabatic hyperspherical potential curves $U_i^+(R)$ for the H_3^+ ion are energy eigenvalues of the fixed- R ionic nuclear motion Schrödinger equation:

$$H_{R=\text{const}} \Phi_i(\theta, \phi; R) = U_i^+(R) \Phi_i(\theta, \phi; R) \quad (2)$$

Here $H_{R=\text{const}}$ is the vibrational hamiltonian, θ and ϕ are two hyperangles, and Φ_i is the eigenfunction. The ionic potential curves are displayed in Fig. 1.

To describe the electron–ion interaction, a set of *ab initio* H_3 potential surfaces and their mutual couplings are first parametrized at fixed values of the nuclear positions using multichannel quantum defect theory (MQDT)^{26,27}. It includes the Jahn–Teller coupling for all principal quantum numbers and the electronic continuum^{16,18,19,28}. The adiabatic hyperspherical potential curves of H_3 are then extracted by searching for autoionizing resonances of the $e\text{--H}_3^+$ system. Each potential curve $U_\beta(R)$ (that is, energy as a function of the hyperradius R) generally exhibits a nonzero autoionization width $\Gamma_\beta(R)$. The initial electron-capture process, followed by the subsequent two-body or three-body fragmentation, can then be described by the DR ‘curve-crossing formalism’²⁹.

Our treatment can be viewed as an application of the vibrational frame transformation of MQDT that has successfully treated diatomic Rydberg state interactions^{26,27}, however, the role of the internuclear distance in diatomic systems is replaced here by the hyperradius R . The quantum-mechanical scattering amplitude for a Jahn–Teller-type symmetry-changing p -wave collision, with the nuclei clamped, will be written as $S_{AA'}(R, \theta, \phi)$. Here (R, θ, ϕ) are the nuclear coordinates, rewritten in terms of a hyperradius R and two hyperangles θ, ϕ . In an earlier study^{18,19} of photoabsorption by a metastable Rydberg state of H_3 , the vibrational frame transformation was employed to obtain the electron–ion scattering matrix, excluding dissociation channels, as a matrix element of $S_{AA'}(R, \theta, \phi)$ connecting two rovibrational eigenfunctions of H_3^+ , integrated over all nuclear coordinates (R, θ, ϕ) . The fixed-nuclei scattering matrix $S_{AA'}(R, \theta, \phi)$ represents an amplitude for the reaction:



where the orbital angular momentum $l = 1$. That is, the electron

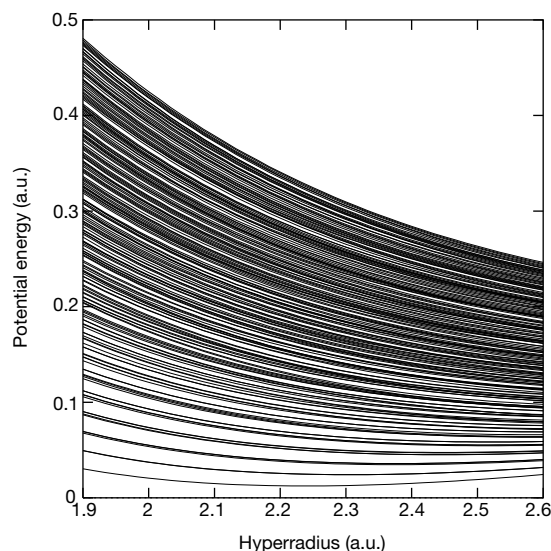


Figure 1 The first 200 adiabatic potential curves of H_3^+ in hyperspherical coordinates. The neglect of mutual coupling between these adiabatic channels is a good approximation here, as shown by the fact that the root-mean-square error of the lowest vibrational levels in the first seven potential curves is 32 cm^{-1} . This error is much smaller than the vibrational spacing of $2,500 \text{ cm}^{-1}$. Note that in the notation $v_1 v_2$, the lowest such potential curve contains the 00^0 vibrational level, and all of the symmetric-stretch vibrational excitations of that level. The energy is in atomic units.

scatters from one irreducible representation (Λ') of the D_{3h} symmetry group into another, Λ , while the nuclei do not have time to move.

For a p -wave electron, the $p\sigma$ -component corresponds to the lone D_{3h} irreducible representation A_1 , while the $p\pi$ -components $\Lambda = \pm 1$ are associated with the Jahn–Teller active modes of irreducible representation E . The Jahn–Teller-induced coupling between the $p\pi$ -components can be represented by a 2×2 reaction matrix^{16,18,19}

$$K = \begin{pmatrix} \delta\rho^2 & \lambda\rho \exp(i\phi) \\ \lambda\rho \exp(-i\phi) & \delta\rho^2 \end{pmatrix} \quad (4)$$

Here (ρ, ϕ) are radial and azimuthal coordinates of the doubly degenerate vibrational modes, and (λ, δ) are the Jahn–Teller coupling constants^{19,28}. In equation (4), the constant, equilibrium geometry quantum defect $\mu^0 = \mu_{\Lambda=\pm 1}$ is not shown. It is included into the treatment in the same way as in ref. 19. The Jahn–Teller and MQDT parameters $(\lambda, \delta, \mu_\Lambda)$ are taken to be independent of the nuclear configuration, which tests show is a good approximation over the range of the vibrational ground state of H_3^+ . We adopt Jahn–Teller parameters determined in ref. 28.

The short-range scattering amplitude for this process is now calculated by a generalized vibrational frame transformation^{26,27} as an integral:

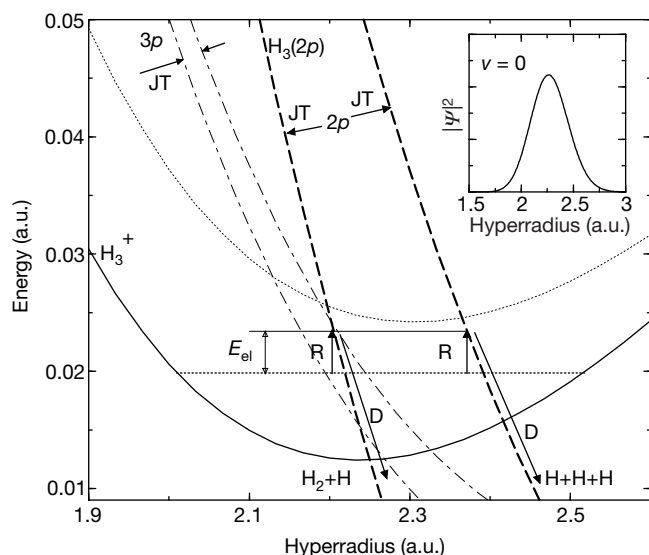


Figure 2 Schematic picture of the hyperspherical representation of the dissociative recombination process. Shown are the dissociative recombination (DR) pathways and the Jahn–Teller splitting of the in-plane $2p$ - and $3p$ -components. The solid curve shows the lowest H_3^+ hyperspherical potential curve, whereas the first excited ionic potential curve is a fine dotted curve. The storage-ring experiments^{1–7} have been performed for vibrationally cold (but rotationally hot) H_3^+ target ions. The energy of this vibrational ground-state level is shown as a horizontal dotted line, and its hyperradial wavefunction is shown as the inset. A solid horizontal line indicates a typical incident electron energy, E_{el} . Two typical Jahn–Teller-split $2p$ components are also shown; these are associated with high ionic hyperspherical potential curves, roughly 50–200 states above the H_3^+ ground-state curve. For each dissociative curve shown, the process of DR is indicated by two arrows, R and D. First, the electron is captured (R), and then the system evolves towards dissociation (D). During the dissociative expansion of the hyperradius, the H_3 resonant state has some probability of autoionizing back to the ionic ground state; if it autoionizes back, it means that DR has not occurred. However, the autoionization widths are small (all of the $\Gamma_\beta(R)$ are less than 2 cm^{-1}), which means that nearly all captures are expected to result in dissociative recombination. In other words, the survival probability is taken to be approximately unity for the ‘upper bound’ calculation. The contributions from alternative channels, namely $H + H + H$ or $H_2(v, j) + H$ can be counted separately, giving partial DR rates or cross-sections.

$$S_{i\Lambda, f\Lambda'}(R) = \langle \langle \Phi_i | S_{\Lambda\Lambda'}(R, \theta, \phi) | \Phi_f \rangle \rangle^{(R)} \quad (5)$$

Here the double brackets denote an integral over the hyperangles only, at a fixed R . The dimension of this S -matrix is $3N \times 3N$, where N is the number of adiabatic eigenstates of H_3^+ that are retained. A standard MQDT closed-channel ‘elimination’, plus a time-delay analysis, determines the resonance positions and partial widths at each hyperradius^{26,27}.

Our treatment differs from that of refs 18 and 19, however, in that we treat the quantization of the vibrating nuclei in two stages, by using hyperspherical coordinate techniques. We deliberately delay the solution of the Schrödinger equation in the hyperradial degree of freedom, in order to gain insight into the relevant DR pathways. This provides an immediate link with the theory of DR in diatomic molecules, with the role of the internuclear distance in diatomics generalized to the hyperradius in polyatomic species. In principle, this generalization will also permit the application of more sophisticated treatments developed for diatomics³⁰ that incorporate the competition between direct and indirect Rydberg state pathways.

In general, each H_3^+ adiabatic potential curve supports an infinite number of autoionizing Rydberg states, at every value of R . The direct contributors to DR are the $2p$ components, which are the only ones that lead diabatically to dissociation. All other higher-excited Rydberg states are energetically closed dissociation channels in this energy range, but nevertheless they can contribute as indirect DR pathways. Once such a Rydberg state is excited during an electron-capture process, the system can oscillate in R many times, and can then undergo a Landau–Zener transition into one of the $2p$ dissociative pathways that are open. The $2p$ symmetry has three different components (with $\Lambda = 0, \pm 1$). The coupled $2p_1$ and $2p_{-1}$ components form two diabatic potential surfaces of H_3 that correlate asymptotically either to $H_2 + H$ or to $H + H + H$. The $2p_0$ state does not lead to a dissociation channel that is open in this sub-eV energy range. Yet an initially captured np_0 electron can be scattered into the $np_{\pm 1}$ states by Coriolis coupling, which gives another indirect pathway to DR although it is a minor contributor to the final rate.

Figure 2 shows how the dissociative recombination process occurs within our framework. To describe the DR process quantitatively, we use the cross-section expression²⁹:

$$\sigma = \sum_{\beta} \frac{\pi^2}{E_{el}} \frac{\Gamma_{\beta}(R_{\beta})}{|U'_{\beta}(R_{\beta})|} |\Psi(R_{\beta})|^2 \quad (6)$$

Here the sum is over all resonant curves crossing the ground-state hyperspherical potential curve for H_3^+ . $\Gamma_{\beta}(R_{\beta})$ is the partial auto-

Table 1 Some of the autoionizing states that contribute to dissociative recombination at an energy E_{el} of 0.025 eV

n_{Ryd}	JT state		N_{ion}	R_{β} (a.u.)	Γ_{β} (cm^{-1})	$ U_{\beta} $ (a.u.)
$ \Psi ^2$						
2	u	53	1.920	0.18	0.325	0.29
3	l	12	1.951	1.10	0.187	0.42
4	u	5	1.957	1.58	0.138	0.45
4	u	6	1.967	0.60	0.133	0.50
5	u	5	2.097	0.74	0.076	1.42
5	l	4	2.091	0.61	0.078	1.37
5	l	5	2.107	0.86	0.073	1.50
5	l	6	2.108	0.62	0.073	1.51
6	u	5	2.245	0.96	0.065	2.21
			2.519	0.45	0.043	0.85
6	l	4	2.234	0.58	0.072	2.20
			2.533	1.03	0.048	0.76
6	l	5	2.277	0.59	0.065	2.22
			2.485	0.50	0.043	1.07
6	l	6	2.477	0.51	0.043	1.12

Letters u and l mean upper and lower states of the Jahn–Teller couple ($\Lambda = \pm 1$). Note that states (6u5), (6l4) and (6l5) cross the horizontal line $E_{el} = 0.025 \text{ eV}$ at two different times. n_{Ryd} is the approximate principal quantum number of the resonance, while N_{ion} is the eigenvalue index for the H_3^+ hyperspherical potential curves, starting from the lowest energy state.

ionization width in the incident channel, E_{cl} is the energy of the capture electron in atomic units, $\Psi(R_\beta)$ is the normalized hyper-radial wavefunction of the lowest H_3^+ vibrational level, shown at the inset of Fig. 2, R_β is the Condon point (that is, the hyperradius at which the β th curve lies at an energy E_{cl} above the H_3^+ ground state energy), and $U'_\beta(R_\beta)$ is the derivative of the excited-state H_3 hyper-spherical potential curve into which capture occurs. For cases where capture occurs into a potential curve that has a minimum near the Condon point, we have implemented a modification to the approximate overlap integral that replaces its contributions in equation (6) by a coherent sum of the values from both Condon points at the lowest vibrational eigenenergy in that potential $(\pi^2/E_{cl})|\Sigma_\beta \Psi(R_\beta)|\Gamma_\beta(R_\beta)/U'_\beta(R_\beta)|^{1/2}|^2$. This modification eliminates the unphysical divergence in the cross-section that would otherwise result at those minima. We have verified that this approximation is reasonable. Table 1 shows the dominant contributing pathways at an incident electron energy of 0.025 eV, and also the leading $2p$ 'direct' contribution. Note that the largest resonance width obtained in our calculations is 1.8 times smaller than the largest width (2.8 cm^{-1}) due to vibrational autoionization that emerges from the theoretical calculations^{18,19,28} and from experiment²⁸. Although this is not fully quantitative agreement, the comparison gives us some

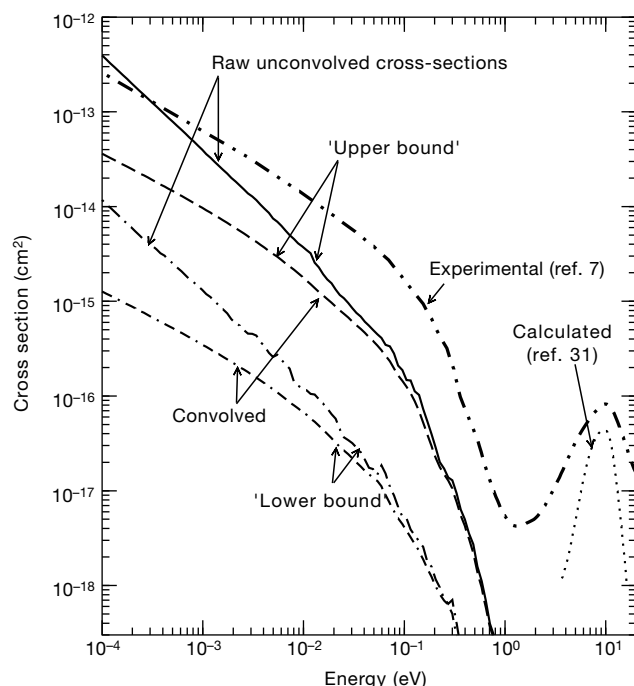


Figure 3 Comparison of experimental and theoretical cross-sections. The solid curve displays the calculated 'upper bound' total cross-section (including the indirect mechanism) as a function of the incident total energy. The dot-dot-dashed curve shows the 'average cross-section' obtained in a recent storage-ring experiment⁷ as a function of the relative longitudinal energy. The theoretical curves to compare with experiment are our convolved 'upper bound' and 'lower bound' calculations, which are discussed in the text. The higher-energy peak in the DR cross-section caused by a doubly excited state, calculated in ref. 31, is also shown for reference. The dot-dashed line shows our unconvolved cross-section for the direct ($2p$ only) DR pathway. Only the longitudinal energy of the electrons is controlled in the experiment, while the transverse energy had a thermal spread of $kT_\perp = 0.025 \text{ eV}$ (ref. 7). The convolved cross-sections have been averaged accordingly, assuming a Maxwell–Boltzmann distribution with a transverse temperature. The resulting theoretical results (plotted versus the parallel energy component) to be compared with experiment are shown as a dashed curve. The unconvolved 'upper-bound' theoretical cross-section includes all np H_3 Rydberg states up to $n = 11$. The raw 'lower-bound' cross-section we obtain from only the direct ($2p$) dissociative mechanism is plotted as a dot-dashed curve. The direct pathway contributes only about 3% of the 'upper-bound' cross-section.

confidence that the electron–ion interaction is described at a reasonable level of approximation.

Figure 3 compares our calculated DR cross-section with the newest storage-ring experiment measurement⁷. The branching ratio has also been calculated, but for the direct $2p$ pathways only; our theoretical results for the fraction of DR events that leads to three-body fragmentation is $\sigma(H + H + H)/\sigma_{DR} = 0.70 \pm 0.07$. A coupled-channel calculation that can track the recombination flux excited by indirect np pathways is beyond the scope of the present study. Consequently, our results only suggest what the final branching ratio might be after the indirect pathways are incorporated more satisfactorily. Nevertheless, it is at least encouraging that the theoretical branching ratio overlaps with the experimental result³ of 0.75 ± 0.08 . Moreover, this direct-pathways calculation results in a peak vibrational distribution for the channels $H_2(v, j)$ that peaks at $v \approx 5-6$, which is in accord with recent experimental results⁴. Here V and j are the vibrational and notational quantum numbers for H_2 .

We have conducted two different calculations of the total recombination rate, in order to assess the role of indirect processes. The first is a 'lower-bound' cross-section that includes only the direct dissociation pathways ($2p$) in the summation of equation (6). The second is an 'upper-bound' cross-section that includes all pathways in equation (6). The latter calculation will be accurate if the coupling between $2p$ and np channels is strong enough for predissociation to occur before the excited state can autoionize—that is, during a few vibrations of the np states. This is plausible because the autoionization lifetimes are so long, of the order of thousands of vibrational periods, but it is far from proved. It will be important for future studies to test this assumption quantitatively by actually mapping out the detailed Landau–Zener crossings between indirect and direct DR pathways. Alternatively, an experiment could help to assess this key issue by measuring the branching ratio between preionization and predissociation of the np Rydberg levels, for n in the range 3–7. We stress that our findings are consistent with previous theoretical work^{14,15,20} which has already suggested and documented a large enhancement of the DR rate through indirect Rydberg state pathways, by as much as two orders of magnitude.

For a thermal distribution of incident electron energies, the DR rate coefficient $\alpha = \langle \nu \sigma \rangle$ is defined as⁵

$$\alpha(kT) = \frac{8\pi}{(2\pi kT)^{3/2}} \int_0^\infty \sigma(E) e^{-E/kT} E dE \quad (7)$$

Table 2 compares our calculations with the results of three recent experiments and with a recent calculation (ref. 14). A more complete comparison with earlier experiments can be found in refs 1 and 8. The 300 K rate constant calculated by ref. 14 is close to our 'lower-bound' rate constant, but this agreement is fortuitous, because our lower-bound value includes only the contributions from the $2p$ 'direct' DR pathways, whereas the result quoted for ref. 14 includes an estimate for the indirect pathways. For that reason, the value from ref. 14 should be compared with our 'upper-bound' result in Table 2; the 'direct' DR rate¹⁴ is another two orders of magnitude smaller.

Some questions still remain to be answered by future studies of the low-energy DR process. We have found that Jahn–Teller coupling is more important by far in the sub-eV energy range

Table 2 Comparison of our calculated rate coefficients with experiments

Reference	α ($\text{cm}^3 \text{s}^{-1}$)
Jensen <i>et al.</i> ⁷ ; storage-ring experiment	1.0×10^{-7} (300 K)
Sundström <i>et al.</i> ⁸ ; storage-ring experiment	1.15×10^{-7} (300 K)
Laubé <i>et al.</i> ⁹ ; FALP-MS experiment	7.8×10^{-8} (300 K)
Glosik <i>et al.</i> ¹⁰ ; AISA experiment	$< 1.3 \times 10^{-8}$ (270 K)
Schneider <i>et al.</i> ¹⁴ ; calculation including indirect paths*	$(3-5) \times 10^{-10}$ (300 K)
Present 'lower bound' calculation including just direct paths	3.4×10^{-10} (300 K)
Present 'upper bound' calculation with indirect paths also	1.2×10^{-8} (300 K)

*Theoretical treatment.

than any previously discussed DR mechanism. Our cross-sections are in fair agreement with the range of FALP measurements but even our upper-bound cross-section remains ten times smaller than the storage-ring experiments that are expected to be highly reliable. The present calculation has not separated out the *para*- and *ortho*-contributions, which simplifies our theoretical description by allowing us to average over all nuclear permutation symmetries without imposing them explicitly.

One candidate process to consider in the future is the set of DR pathways excited by the inverse of rotational autoionization. Those pathways require a quantized treatment of the ionic rotational channels, which has not yet been incorporated into our present calculations. Rotationally autoionizing *np* resonances are known to have some of the largest widths observed for H_3 Rydberg states^{17–19,28}, but they do not necessarily generate a high DR contribution because these states probably autoionize back to the electron continuum before they have time to dissociate. An investigation of this pathway may help to understand why current generation storage-ring experiments with rotationally hot target ions appear to have a much larger DR rate than the FALP experiments. For example, in ref. 4, $kT_{\text{rot}} = 0.23$ eV, which implies a peak in the rotational population at the $J^+ = 5$ state. For target ions in the $J^+ = 5$ incident channel, electron capture can occur via inverse rotational autoionization at energies up to the $J^+ = 7$ ionic threshold, which is about 0.13 eV.

Although the present study has not solved the problem of why different measurements of the low-energy H_3 DR rate give such different results, our hope is that this identification of the dominant DR mechanism will stimulate further progress in both experiment and theory in the near future. □

Received 4 April; accepted 4 July 2001.

- Larsson, M. Experimental studies of the dissociative recombination of H_3^+ . *Phil. Trans. R. Soc. Lond. A* **358**, 2433–2443 (2000).
- Larsson, M. *et al.* Direct high-energy neutral-channel dissociative recombination of cold H_3^+ in an ion storage ring. *Phys. Rev. Lett.* **70**, 430–433 (1993).
- Datz, S. *et al.* Branching-processes in the dissociative recombination of H_3^+ . *Phys. Rev. Lett.* **74**, 896–899 (1995).
- Strasser, D. *et al.* Two- and three-body kinematical correlation in the dissociative recombination of H_3^+ . *Phys. Rev. Lett.* **86**, 779–782 (2001).
- Sundström, G. *et al.* Destruction rate of H_3^+ by low-energy electrons measured in a storage-ring experiment. *Science* **263**, 785–787 (1994).
- Tanabe, T. *et al.* in *Dissociative Recombination: Theory, Experiment and Applications* Vol. IV (eds Larsson, M. Mitchell, J. B. A. & Schneider, I. F.) 170–179 (World Scientific, Singapore, 2000).
- Jensen, M. J. *et al.* Dissociative recombination and excitation of H_3^+ . *Phys. Rev. A* **63**, 052701–1–052701-5 (2001).
- Laubé, S. *et al.* New FALP-MS measurements of H_3^+ , D_3^+ and HCO^+ dissociative recombination. *J. Phys. B* **31**, 2111–2128 (1998).
- Glosik, J., Plasil, R., Poterya, V., Kurdna, P. & Tichý, M. The recombination of H_3^+ ions with electrons: dependence on partial pressure of H_2 . *Chem. Phys. Lett.* **331**, 209–214 (2000).
- Amano, T. The dissociative recombination rate coefficients of H_3^+ , D_3^+ , and HCO^+ . *J. Chem. Phys.* **92**, 6492–6501 (1990).
- McCall, B. J., Geballe, T. R., Hinkle, K. H. & Oka, T. Detection of H_3^+ in the diffuse interstellar medium toward Cygnus OB2 No. 12. *Science* **279**, 1910–1913 (1998).
- Oka, T. Astronomy, physics and chemistry of H_3^+ —Introductory remarks. *Phil. Trans. R. Soc. Lond. A* **358**, 2363–2369 (2000).
- Geballe, T. R. H_3^+ between the stars. *Phil. Trans. R. Soc. Lond. A* **358**, 2503–2512 (2000).
- Schneider, I. F., Orel, A. E. & Suzor-Weiner, A. Channel mixing effects in the dissociative recombination of H_3^+ with slow electrons. *Phys. Rev. Lett.* **85**, 3785–3788 (2000).
- Orel, A. E., Schneider, I. F. & Suzor-Weiner, A. Dissociative recombination of H_3^+ : progress in theory. *Phil. Trans. R. Soc. Lond. A* **358**, 3293–3293 (2000).
- Staib, A. & Domcke, W. Analysis of the Jahn–Teller effect in the np^2E' Rydberg series of H_3 and D_3 . *Z. Phys. D* **16**, 275–282 (1990).
- Bordas, M. C., Lembo, L. J. & Helm, H. Spectroscopy and multichannel quantum-defect theory analysis of the *np* Rydberg series of H_3 . *Phys. Rev. A* **44**, 1817–1827 (1991).
- Stephens, J. A. & Greene, C. H. Quantum-defect description of H_3 Rydberg state dynamics. *Phys. Rev. Lett.* **72**, 1624–1627 (1994).
- Stephens, J. A. & Greene, C. H. Quantum-defect description of H_3 and the Jahn–Teller effect. *J. Chem. Phys.* **102**, 1579–1591 (1995).
- Bates, D. R. Dissociative recombination when potential energy curves do not cross. *J. Phys. B* **25**, 5479–5488 (1992).
- Cecchi-Pestellini, C. & Dalgarno, A. H_3^+ in diffuse interstellar gas. *Mon. Not. R. Astron. Soc.* **313**, L6–L8 (2000).
- Jaquet, R., Cencek, W., Kutzelnigg, W. & Rychlewski, J. Sub-microhertz accuracy potential energy surface for H_3^+ including adiabatic and relativistic effects II. Rovibrational analysis for H_3^+ and D_3^+ . *J. Chem. Phys.* **108**, 2837–2846 (1998).

- Esry, B. D., Lin, C. D. & Greene, C. H. Adiabatic hyperspherical study of the helium trimer. *Phys. Rev. A* **54**, 394–401 (1996).
- Lin, C. D. Hyperspherical coordinate approach to atomic and other coulombic 3-body systems. *Phys. Rep.* **257**, 2–83 (1995).
- Zhou, Y., Lin, C. D. & Shertzer, J. Hyperspherical approach to coulombic 3-body systems with different masses. *J. Phys. B* **26**, 3937–3949 (1993).
- Greene, C. H. & Jungen, Ch. Molecular applications of quantum defect theory. *Adv. At. Mol. Phys.* **21**, 51–121 (1985).
- Jungen, Ch. *Molecular Applications of Quantum Defect Theory* (Institute of Physics, Bristol, 1996).
- Mistrik, I. *et al.* Ab initio analysis of autoionization of H_3 molecules using multichannel quantum-defect theory and new quantum-defect surfaces. *Phys. Rev. A* **61**, 033410–1–033410-16 (2000).
- O'Malley, T. F. Theory of dissociative attachment. *Phys. Rev.* **150**, 14–29 (1966).
- Giusti, A. A multichannel quantum defect approach to dissociative recombination. *J. Phys. B* **13**, 3867–1894 (1980).
- Orel, A. E. & Kulander, K. C. Resonant dissociative recombination of H_3^+ . *Phys. Rev. Lett.* **71**, 4315–4318 (1993).

Acknowledgements

We thank J. A. Stephens, A. E. Orel and A. Suzor-Weiner for discussions, and M. Larsson, G. Dunn, S. Datz, T. Oka and A. Dalgarno for comments on the experimental and astrophysical issues. Computational work was carried out at the National Energy Research Supercomputer Center at Lawrence Berkeley National Laboratory. This work was supported by the National Science Foundation.

Correspondence and requests for materials should be addressed to C.H.G. (e-mail: chris.greene@colorado.edu).

Parasitic computing

Albert-László Barabási*, Vincent W. Freeh†, Hawoong Jeong* & Jay B. Brockman†

* Department of Physics; and † Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, Indiana 46556, USA

Reliable communication on the Internet is guaranteed by a standard set of protocols, used by all computers¹. Here we show that these protocols can be exploited to compute with the communication infrastructure, transforming the Internet into a distributed computer in which servers unwittingly perform computation on behalf of a remote node. In this model, which we call ‘parasitic computing’, one machine forces target computers to solve a piece of a complex computational problem merely by engaging them in standard communication. Consequently, the target computers are unaware that they have performed computation for the benefit of a commanding node. As experimental evidence of the principle of parasitic computing, we harness the power of several web servers across the globe, which—unknown to them—work together to solve an NP complete problem².

Sending a message through the Internet is a sophisticated process regulated by layers of complex protocols. For example, when a user selects a URL (uniform resource locator), requesting a web page, the browser opens a transmission control protocol (TCP) connection to a web server. It then issues a hyper-text transmission protocol (HTTP) request over the TCP connection. The TCP message is carried via the Internet protocol (IP), which might break the message into several packages, that navigate independently through numerous routers between source and destination. When an HTTP request reaches its target web server, a response is returned via the same TCP connection to the user's browser. The original message is reconstructed through a series of consecutive steps, involving IP and TCP; it is finally interpreted at the HTTP level, eliciting the appropriate response (such as sending the requested web page)¹. Thus, even a seemingly simple request for a web page involves a significant amount of computation in the network and at the computers at the end points. The success of the Internet rests on the cooperation of and trust between all involved parties.

Here we demonstrate that the trust-based relationships between machines connected on the Internet can be exploited to use the resources of multiple servers to solve a problem of interest without authorization. In essence, a 'parasitic computer' is a realization of an abstract machine for a distributed computer that is built upon standard Internet communication protocols. We use a parasitic computer to solve the well known NP-complete satisfiability problem, by engaging various web servers physically located in North America, Europe, and Asia, each of which unknowingly participated in the experiment. Like the SETI@home project (see <http://www.seti.org>), parasitic computing decomposes a complex problem into computations that can be evaluated independently and solved by computers connected to the Internet; unlike the SETI project, however, it does so without the knowledge of the participating servers. Unlike 'cracking' (breaking into a computer) or computer viruses, however, parasitic computing does not compromise the security of the targeted servers, and accesses only those parts of the servers that have been made explicitly available for Internet communication.

To solve many NP complete problems, such as the travelling salesman or the satisfiability problem, a common technique is to generate a large number of candidate solutions and then test the candidates for their adequacy. Because the candidate solutions can be tested in parallel, an effective computer architecture for these problems is one that supports simultaneous evaluation of many tests. An example of such a machine is illustrated in Fig. 1a. Here, the computer consists of a collection of target nodes connected to a network, where each of the target nodes contains an arithmetic and logic unit (ALU) that is capable of performing the desired test and a network interface (NIF) that allows the node to send and receive messages across the network. A single home parasite node initiates the computation, sends messages to the targets directing them to perform the tests, and tabulates the results.

Owing to the many layers of computation involved in receiving and interpreting a message, there are several Internet protocols that, in principle, could be exploited to perform a parasitic computation. For example, an IP-level interpretation could force routers to solve the problem, but such an implementation creates unwanted local bottlenecks. To truly delegate the computation to a remote target computer, we need to implement it at the TCP or higher levels. Potential candidate protocols include TCP, HTTP, or encryption/decryption with secure socket layer (SSL).

During package transfer across the Internet, messages can be corrupted, that is, the bits can change. TCP contains a checksum that provides some data integrity of the message. To achieve this, the sender computes a checksum and transmits that with the message. The receiver also computes a checksum, and if it does not agree with the sender's, then the message was corrupted (see Fig. 2). One property of the TCP checksum function is that it forms a sufficient logical basis for implementing any Boolean logic function, and by extension, any arithmetic operation³.

To implement a parasitic computer using the checksum function we need to design a special message that coerces a target server into performing the desired computation. As a test problem we choose to solve the well known 'satisfiability' (or SAT) problem, which is a common test for many unusual computation methods^{4,5}. The SAT problem involves finding a solution to a Boolean equation that satisfies a number of logical clauses. For example, $(x_1 \text{ XOR } x_2) \text{ AND } (\sim x_2 \text{ AND } x_3)$ in principle has 2^3 potential solutions, but it is satisfied only by the solution $x_1 = 1$, $x_2 = 0$, and $x_3 = 1$. This is called a 2-SAT problem because each clause, shown in parentheses, involves two variables. The more difficult 3-SAT problem is known to be NP complete, which in practice means that there is no known polynomial-time algorithm which solves it. Indeed, the best known algorithm for an n -variable SAT problem scales exponentially with n , following $(1.33)^n$ (ref. 6). Here we follow a brute-force approach by instructing target computers to evaluate, in a distributed fashion, each of the 2^n potential solutions.

A possible parallel approach to the SAT problem is shown schematically in Fig. 1. The parasite node creates 2^n specially constructed messages designed to evaluate a potential solution. The design of the message, exploiting the TCP checksum, is described in Fig. 3. These messages are sent to many target servers throughout the Internet. Note that while we choose the simpler 2-SAT case to illustrate the principle behind the coding, the method can be extended to code the NP complete 3-SAT problem as well, as explained in the Supplementary Information (see also <http://www.nd.edu/~parasite>). The message received by a target server contains an IP header, a TCP header, and a candidate solution (values for x_i). The operators in the Boolean equation determine the value of the checksum, which is in the TCP header. The parasite node injects each message into the network at the IP level (Fig. 1), bypassing TCP. After receiving the message, the target server verifies the data integrity of the TCP segment by calculating a TCP

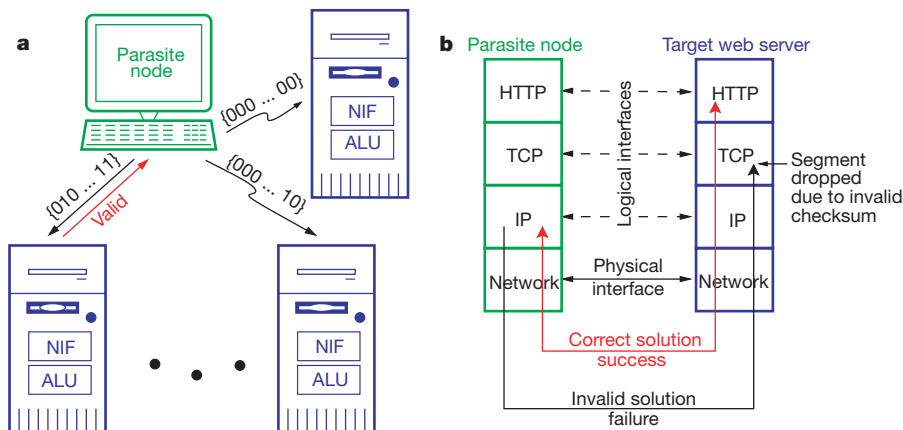


Figure 1 Schematic diagram of our prototype parasitic computer. **a**, A single parasite node (green) coordinates the computations occurring remotely in the Internet protocols. It sends specially constructed messages to some number of targeted nodes (blue boxes), which are web servers consisting of an arithmetic and logic unit (ALU) and a network interface (NIF). **b**, Levels of communication between the parasitic node and the target in our proof-of-principle implementation. The packet carrying the problem to be solved is inserted into the network at the IP level, bypassing the parasite node's TCP. The

construction of the message is such that an invalid solution fails the TCP checksum and is dropped, which is illustrated by the black path labelled 'failure'. Consequently, only valid solutions propagate up through the TCP protocol layer to HTTP, illustrated by the red path labelled 'success'. Targeted web servers respond to all requests that reach HTTP, even invalid HTTP requests. Thus, the parasite node sends out a message for each possible solution (black arrow in **a**), but only receives a response for each solution that is valid (red arrow in **a**).

checksum. The construction of the message (Fig. 3) ensures that the TCP checksum fails for all messages containing an invalid solution to the posed SAT problem. Thus, a message that passes the TCP checksum contains a correct solution. The target server will respond to each message it receives (even if it does not understand the request). As a result, all messages containing invalid solutions are dropped in the TCP layer. Only a message which encodes a valid solution 'reaches' the target server, which sends a response to the 'request' it received.

We have implemented the above scheme using as a parasitic node an ordinary desktop machine with TCP/IP networking. The targeted computers are various web servers physically located in North America, Europe, and Asia, each of which unwittingly participated in the experiment. As explained in Fig. 1, our parasite node distributed 2^n messages between the targets. Because only messages containing valid solutions to the SAT problem pass through TCP, the target web server received only valid solutions. This is interpreted as an HTTP request, but it is of course meaningless in this context. As required by HTTP, the target web server sends a response to the parasitic node, indicating that it did not understand the request. The parasite node interprets this response as attesting to the validity of the solution. As expected and by design, incorrect

solutions do not generate responses for the web server. A typical message sent by the parasite, and a typical response from a target web server are included in the Supplementary Information.

Our technique does not receive a positive acknowledgement that a solution is invalid because an invalid solution is dropped by TCP. Consequently, there is a possibility of false negatives: cases in which a correct solution is not returned, which can occur for two reasons. First, the IP packet could be dropped, which might be due to data corruption or congestion. Normally TCP provides a reliability mechanism against such events⁷, but our current implementation cannot take advantage of this. Second, because this technique exploits the TCP checksum, it circumvents the function the checksum provides. The TCP checksum catches errors that are not caught in the checks provided by the transport layer, such as errors in

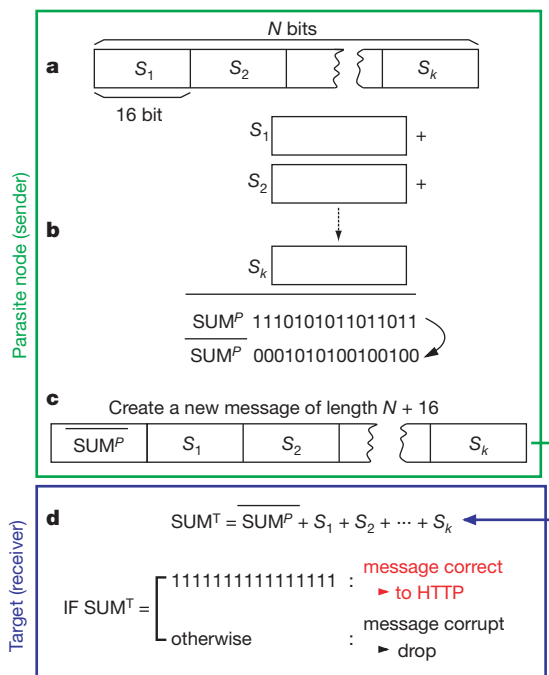


Figure 2 TCP checksum. **a**, The checksum is a simple function performed by all web servers (in TCP), which allows a recipient to check if the received message has been corrupted during transmission. The sender (parasitic node) breaks the message consisting of N bits into 16-bit words, shown as $S_1, S_2, \dots, S_k$. **b**, The k words are added together using binary one's-complement arithmetic, providing the sum denoted as SUM^P . Next, the sender performs a bit-wise complement on the checksum, so that every bit is flipped: a 0 becomes a 1 and a 1 becomes a 0, obtaining $SUM^{\bar{P}}$. An example of a checksum and its complement are shown in the figure. **c**, The sender incorporates the complement of the checksum into the header of the message. **d**, The receiving computer (target) again breaks the received message into 16-bit segments and adds them together. The value of the checksum SUM^T calculated by the target is $SUM^{\bar{P}} + SUM^P$, the first term coming from the header and the second term being the contribution from $S_1 + S_2 + \dots + S_k$. As $SUM^{\bar{P}}$ and SUM^P are complementary, the checksum obtained by the receiver has to be 1111111111111111. If any bit along the message has been corrupted during transmission, the checksum obtained by the target will be different from all ones, in which case the target drops the message. A non-corrupted message is passed to the HTTP protocol, which will attempt to interpret its content.

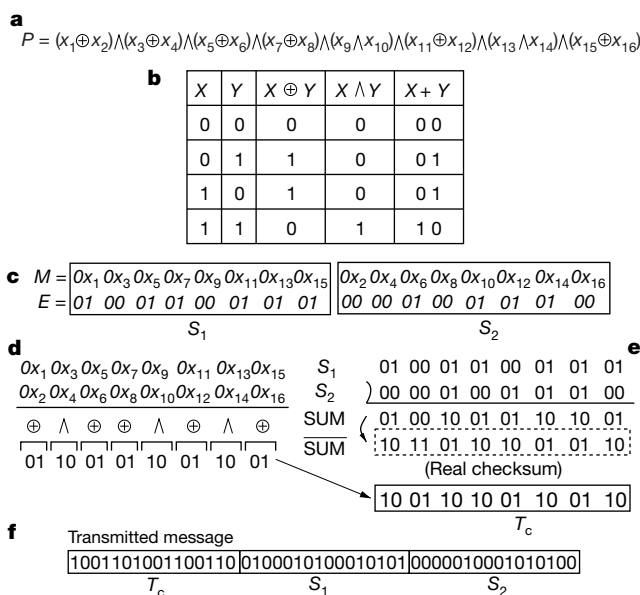


Figure 3 Deciding satisfiability using checksum. The 2-SAT problem shown in **a** involves 16 variables $\{x_1, x_2, \dots, x_{16}\}$ and the operators AND ($\wedge$) and XOR ($\oplus$). **b**, The logical table of XOR, AND and the binary sum (+). In order to get a TRUE answer for P , each clause shown in separate parentheses in **a** needs to be independently TRUE. **c**, To evaluate the value of P we generate a 32-bit message M that contains all 16 variables, each preceded by a zero. As an illustration, we show a possible solution E . TCP groups the bits of the received message in two 16-bit segment and adds them together (Fig. 2a). As shown in **d**, this will result in adding each (x_i, x_{i+1}) pair together where i is odd. The sum can have four outcomes. Comparing the sum with the ($\oplus$) column in the table in **b**, we notice that a TRUE answer for the XOR clause $(x_i \oplus x_{i+1})$ coincides with the (01) result of the (x_i, x_{i+1}) sum. Similarly, if the clause has an AND operator, $(x_i \wedge x_{i+1})$ is true only when the checksum is (10). This implies that for a set of variables $\{x_1, x_2, \dots, x_{16}\}$ that satisfies P the checksum will be determined by the corresponding operators only (that is, a $\oplus$ should give (01) for the sum check, and for $\wedge$ the sum is (10)). For illustration, in **d** we show the formal lineup of the variables, while in **e** we show an explicit example. The correct complemented checksum for E should be $\bar{SUM} = 10110110100110$. In contrast, the parasitic computer places in the header the transmitted checksum $T_c = 1001101001100110$, which is uniquely determined by the operators in P , as shown in **d**. To turn the package into a parasitic message the parasitic node prepares a package, shown in **f**, preceded by a checksum T_c , and continued by a 32-bit sequence (S_1, S_2), which represent one of the 2^{16} potential solutions. If S_1 and S_2 do not represent the correct solution, then the checksum evaluated by the target TCP will not give the correct sum (111...11). The TCP layer at the target concludes that the message has been corrupted, and drops it. However, if S_1 and S_2 contain the valid solution, the message is sent to HTTP. The web server interprets the solution as an HTTP request; however, because it is not a valid HTTP request, the web server responds with a message saying something like 'page not found' (red arrow in Fig. 1a). Thus, every message to which the parasite node receives a response is a solution to the posed SAT problem (see Fig. 2d).

intermediate routers and the end points⁸. Measurements show that the TCP checksum fails in about 1 in 2^{10} messages⁹. The actual number of TCP checksum failures depends on the communication path, message data, and other factors. To test the reliability of our scheme, we repeatedly sent the correct solution to several host computers located on three continents. The rate of false negatives with our system ranged from 1 in about 100 to less than 1 in 17,000.

The implementation offered above represents only a proof of concept of parasitic computing. As such, our solution merely serves to illustrate the idea behind parasitic computing, and it is not efficient for practical purposes in its current form. Indeed, the TCP checksum provides a series of additions and a comparison at the cost of hundreds of machine cycles to send and receive messages, which makes it computationally inefficient. To make the model viable, the computation-to-communication ratio must increase until the computation exported by the parasitic node is larger than the amount of cycles required by the node to solve the problem itself instead of sending it to the target. However, we emphasize that these are drawbacks of the presented implementation and do not represent fundamental obstacles for parasitic computing. It remains to be seen, however, whether a high-level implementation of a parasitic computer, perhaps exploiting HTTP or encryption/decryption could execute in an efficient manner.

It is important to note that parasitic computing is conceptually related but philosophically different for cluster computing¹⁰, which links computers such that their cumulative power offers computational environments comparable to the best supercomputers. A prominent example of cluster computing is the SETI program, which has so far enlisted over 2.9 million computers to analyse radio signals in search of extraterrestrial intelligence. In cluster computing, the computer's owner willingly downloads and executes software, which turns his computer into a node of a vast distributed computer. Thus, a crucial requirement of all cluster computing models is the cooperation of the computer's owner. This is also one of its main limitations, as only a tiny fraction of computer owners choose to participate in such computations. In this respect, parasitic computing represents an ethically challenging alternative for cluster computing, as it uses resources without the consent of the computer's owner.

Although parasitic computing does not compromise the security of the target, it could delay the services the target computer normally performs, which would be similar to a denial-of-service attack, disrupting Internet service^{11,12}. Thus, parasitic computing raises interesting ethical and legal questions regarding the use of a remote host without consent, challenging us to think about the ownership of resources made available on the Internet. Because parasitic computation exploits basic Internet protocols, it is technically impossible to stop a user from launching it. For example, changing or disrupting the functions that are exploited by parasitic computing would simply eliminate the target's ability to communicate with the rest of the Internet.

In summary, parasitic computing moves computation onto what is logically the communication infrastructure of the Internet, blurring the distinction between computing and communication. We have shown that the current Internet infrastructure permits one computer to instruct other computers to perform computational tasks that are beyond the target's immediate scope. Enabling all computers to swap information and services they are needed could lead to unparalleled emergent behaviour, drastically altering the current use of the Internet^{13,14}. □

Received 21 May; accepted 6 July 2001.

- Peterson, L. L. & Davie, B. S. *Computer Networks, A Systems Approach* 2nd edn (Morgan Kaufmann, San Francisco, 2000).
- Garey, M. & Johnson, D. S. *Computers and Intractability: A Guide to the Theory of NP-completeness* (Freeman, San Francisco, 1979).
- Boole, G. *An Investigation of the Laws of Thought, on Which Are Founded the Mathematical Theories of Logic and Probabilities* (Dover Advanced Mathematics, Macmillan, London, 1854).

- Alderman, L. M. Molecular computation of solutions to combinatorial problems. *Science* **266**, 1021–1024 (1994).
- Ouyang, Q., Kaplan, P. D., Liu, S. & Libchaber, A. DNA solution of the maximal clique problem. *Science* **278**, 446–449 (1997).
- Schöningh, U. in *Proc. 40th Annu. IEEE Conf. Found. Comp. Sci. (FOCS)* 410–414 (IEEE Comp. Sci., Los Alamitos, California, 1999).
- Stevens, W. R. *TCP/IP Illustrated* 144–147 (Addison-Wesley, Reading, Massachusetts, 1994).
- Stone, J. & Partridge, C. in *Proc. ACM SIGCOMM* 309–319 (2000).
- Stone, J., Greenwald, M., Partridge, C. & Hughes, J. Performance of checksums and CRCs over real data. *IEEE Trans. Networking* **6**, 529–543 (1998).
- Foster, I. Internet computing and the emerging grid. *Nature Web Matters* at <http://www.nature.com/nature/webmasters/grid/grid.html> (7 Dec. 2000).
- Cohen, R., Erez, K., ben-Avraham, D. & Havlin, S. Resilience of the Internet to random breakdowns. *Phys. Rev. Lett.* **85**, 4626–4629 (2000).
- Cohen, R., Erez, K., ben-Avraham, D. & Havlin, S. Breakdown of the Internet under intentional attack. *Phys. Rev. Lett.* **86**, 3682–3685 (2001).
- Lawrence, S. & Giles, C. L. Accessibility of information on the web. *Nature* **400**, 107–109 (1999).
- Lawrence, S. & Giles, C. L. Searching the World Wide Web. *Science* **280**, 98 (1998).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank M. Crovella, G. W. Flake, L. Giles, S. Lawrence, D. Peenock and L. Peterson for comments on the manuscript. This work was supported by the NSF.

Correspondence and requests for materials should be addressed to A.-L.B. (e-mail: alb@nd.edu).

Universal behaviour in compressive failure of brittle materials

Carl E. Renshaw* & Erland M. Schulson†

* Department of Earth Sciences; and † Thayer School of Engineering, Dartmouth College, Hanover, Hampshire 03755, USA

Brittle failure limits the compressive strength of rock and ice when rapidly loaded under low to moderate confinement. Higher confinement or slower loading results in ductile failure once the brittle–ductile transition is crossed. Brittle failure begins when primary cracks initiate and slide, creating wing cracks at their tips^{1–3}. Under little to no confinement, wing cracks extend and link together, splitting the material into slender columns which then fail. Under low to moderate confinement, wing crack growth is restricted and terminal failure is controlled by the localization of damage along a narrow band. Early investigations proposed that localization results from either the linkage of wing cracks^{1–3} or the buckling of microcolumns created between adjacent wing cracks^{4,5}. Observations of compressive failure in ice⁶ suggest a mechanism whereby localization initiates owing to the bending-induced failure of slender microcolumns created between sets of secondary cracks emanating from one side of a primary crack. Here we analyse this mechanism, and show that it leads to a closed-form, quantitative model that depends only on independently measurable mechanical parameters. Our model predictions for both the brittle compressive strength and the brittle–ductile transition are consistent with data from a variety of crystalline materials, offering quantitative evidence for universal processes in brittle failure and for the broad applicability of the model.

Beginning at 20–30% of the terminal failure stress, high-resolution images of brittle compressive failure in ice⁶ reveal both wing cracks and secondary cracks emanating from one side of a sliding primary crack (Fig. 1). The secondary cracks create parallel sets of slender microcolumns which are fixed on one end and free on the other. In contrast to previous failure models which have invoked microcolumns as a key element in the failure process^{3–5}, the

observed microcolumns form along a single primary crack and have one free and one fixed end rather than forming between the wing cracks of two adjacent primary cracks and having two fixed ends. We suggest that the onset of the localization of failure, which limits compressive strength, is triggered by the bending-induced failure of these fixed-free microcolumns. The microcolumns bend and fail owing to frictional drag along their free ends, caused by sliding along the primary crack, rather like the breaking of teeth in a comb under a sliding thumb. Once triggered, we assume that the incipient zone of localized damage is sufficient to focus additional microcracking ahead of itself in the manner of a process zone⁷, although we have not attempted to quantitatively model this aspect of the failure process. As the zone advances, it creates, we imagine, a discrete band of intense damage inclined by $\sim 30^\circ$ to the direction of highest compressive stress that becomes the macroscopic fault.

To quantify this process, consider an optimally oriented primary crack inclined at 45° to the most compressive stress σ_1 (Fig. 1). As compression increases, frictional sliding along the primary crack will occur if the confinement ratio $R = \sigma_3/\sigma_1 < (1 - \mu)/(1 + \mu)$, where μ is the coefficient of friction along the sliding segment of the crack and σ_3 is the least compressive principal stress. Higher ratios suppress frictional sliding and lead to ductile failure. Even when R is less than this value, ductile behaviour will occur, we suggest, when the applied strain rate is sufficiently low to allow creep deformation to relax stress concentrations along the primary crack. Then microcolumns no longer form and thus are not available to trigger the localization of damage.

Sliding along the primary crack generates a stress concentration at the end of the sliding segment. This stress concentration results in a region around the sliding segment tip where the stress exceeds the yield stress of the material, resulting in inelastic deformation. If the applied strain rate is high enough, creep deformation will not have sufficient time to relax the stress concentration and a secondary crack will initiate. Secondary cracks do not initiate at lower strain rates because the stress concentration is relaxed by creep. Creep relaxation is governed by a power-law relationship $\dot{\epsilon} = B\sigma^n$, where $\dot{\epsilon}$ is the creep rate caused by the driving stress σ , B is a temperature and material dependent constant, and n is a material constant. We suggest that the brittle-ductile transition occurs when the size of the region of creep deformation just equals the size of the region where the stress exceeds the yield stress of the material. The creep zone size is estimated from the Riedel-Rice⁸ model of creep zones around non-interacting cracks⁹ and the elastic zone estimated using Irwin's

approximation¹⁰. Accordingly, the brittle-ductile transitional strain rate $\dot{\epsilon}_t$ can be expressed in non-dimensional form as

$$\dot{\epsilon}_D = \frac{\dot{\epsilon}_t c^{1.5}}{BK_{Ic}^3} \approx \frac{8}{\sqrt{\pi[(1-R) - \mu(1+R)]}} \quad \text{where } R < \frac{1-\mu}{1+\mu} \quad (1)$$

Here $2c$ is the primary crack size, which in laboratory samples is set by the dimensions of the grain boundary facets, and K_{Ic} is the critical stress intensity factor (Table 1). Details on the derivation of this expression are given in Part A of the Supplementary Information. Implicit in the formulation is $n = 3$, a commonly observed value. Figure 2 demonstrates the excellent agreement between the predictions of equation (1) and the deformation behaviour versus confinement ratio of laboratory specimens of numerous crystalline solids (ice, rock and silicate minerals). If the variation of the parameters B and K_{Ic} as a function of temperature are known, a similar prediction of the transitional strain rate as a function of temperature is possible. For example, between -40° and 0°C , the parameter B for freshwater ice increases by a factor of about 200 (ref. 11), while K_{Ic} decreases by a factor of two¹². The predicted increase in transitional strain rate is in reasonable agreement with the order of magnitude increase observed in the laboratory¹¹. Unfortunately, the data are insufficient to similarly test model predictions for rock.

Next we consider the brittle compressive strength. Equation (1) defines the limits of the confinement ratio and strain rate required for the initiation of secondary cracks, which we consider to be the critical step in the sequence of events leading to brittle failure. Once secondary cracks exist, terminal failure is governed by the load required to trigger the bending-induced failure of the slender microcolumns created between them. Failure of a microcolumn occurs when the stress concentration at the tip of a secondary crack results in its out-of-plane extension. The initiation of out-of-plane extension is related to the compressive loading along the columns, the net bending moment imparted by frictional drag along the free end of the microcolumn, and the microcolumn geometry. In previous work we have suggested that this situation is similar to the out-of-plane extension of an edge crack in a brittle plate^{6,13}. Here we extend this earlier model to include the effects of friction between columns. In this case the failure stress σ_f can be approximated, in dimensionless form, as

$$\sigma_D = \frac{\sigma_f \sqrt{c}}{K_{Ic}} = \frac{2}{\left[\left(1 + \left(1 - \mu \frac{1+R}{1-R} \right)^{2/3} \right)^{1/2} - 1 \right]^{1/2} (1 + 3\mu^2 \alpha^2 (1-R)^2)^{1/2}} \quad (2)$$

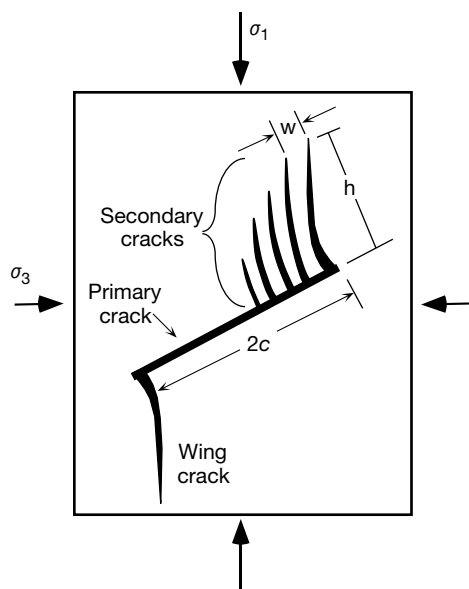


Figure 1 Microcrack geometry.

Table 1 Material properties

Material	$2c$ (mm)	K_{Ic} (MPa m ^{1/2})	Creep-constant (MPa ⁻³ s ⁻¹)*	Temperature (K)
Freshwater ice ²⁰	1–4	0.1	4.3×10^{-7}	263
Saltwater ice ²¹	1–4	0.1	5.1×10^{-6}	263
Clinopyroxene ²²	0.1	2.0	$\log(A) = 8.04, Q = 380$	873
Amphibolite ²³	0.1	2.0	$\log(A) = -1.60, Q = 234$	923–1223
Garnet ²⁴	0.1	1.0	$\log(A) = 17.42, G = 8.5–11, g = 32, T_m = 1,510–1,670$	1223–1528
Sandstone ²⁵	1.0	1.0	2.63×10^{-12}	298
Limestone and dolomite ^{15,26}	0.1	1.0	$\log(A) = -11.5, Q = 4$	298–873
Marble ²²	0.1	2.0	$\log(A) = 31.77, Q = 259$	298
Granite ^{26–29}	0.1	2.0	$\log(A) = -5.7, Q = 187$	298–1173
Gneiss ³⁰	0.1	2.0	$\log(A) = -5.7, Q = 187$	298

* If data were collected at a single temperature, only the creep constant B is given. If data were collected at multiple temperatures, creep coefficients are given and the creep constant may be calculated using: (1) $B = A \exp(-Q/RT)$, where A is an experimental constant (in MPa⁻³ s⁻¹), Q is the activation energy (kJ mol⁻¹), R is the universal gas constant and T the temperature; or (2) $B = AG^{-3} \exp(gT_m/T)$, where A (in s⁻¹) and g are experimental constants, G is the shear modulus (in GPa), and T_m is the melting temperature (in K).

where $\alpha = h/w$ is the aspect, or slenderness, ratio of the microcolumns. The more slender the microcolumns, the weaker the material. A complete derivation of this expression is provided in Part B of the Supplementary Information.

The geometry of the microcolumns depends on the length and spacing of the secondary cracks. For $R > 0$, the extension of the secondary cracks is stable, with the length asymptotically approaching the limiting value³

$$\frac{h}{c} = \frac{1 - R - \mu(1 + R)}{4.3R} \quad (3)$$

Experimental data for ice¹⁴ and rock¹⁵ suggest that axial splitting is suppressed when the length of the secondary cracks is less than the size of the primary crack $h \leq 2c$, implying that $R \geq 0.05$ for shear faulting.

The column width is set by the spacing of the secondary cracks. This, we suggest, is related to the extent of the shear stress concentration just ahead of a sliding segment along the primary crack. The sliding segment of the primary crack is defined by the region of the primary crack that has already separated (fully decohered) and along which sliding is governed by friction. Ahead of the sliding segment, the primary crack has not separated, but sliding may occur by a viscous process that results in partial decohesion¹⁶. The rate of viscous sliding and decohesion immediately ahead of the sliding segment is accelerated because the shear

stress is intensified there. As a result, non-uniform sliding occurs along the primary crack which induces tensile stress within material on one side of the boundary¹⁷ and results in the nucleation of a secondary crack if the applied strain rate is above the brittle–ductile transition. At the same time, the accelerated viscous sliding ahead of the sliding segment weakens the primary crack over a distance equal to the extent of the zone of shear stress concentration, eventually resulting in the extension of the sliding segment into this region. This process repeats itself, with the sliding segment extending, nucleating a secondary crack, and then extending again, culminating in a set of secondary cracks. Their spacing is controlled by the extent of the zone of shear stress concentration ahead of the sliding segment. Here we approximate the boundary of this zone as the location where the magnitude of the gradient of the normalized shear stress induced by a mode II crack¹⁸ exceeds unity, yielding

$$\frac{w}{c} = \left[1 + \left(1 - \mu \frac{1 + R}{1 - R} \right)^{2/3} \right]^{1/2} - 1 \quad (4)$$

The normalizing shear stress is the shear stress that would exist in the absence of any cracks. The complete derivation is provided in Part B of the Supplementary Information.

Figure 3 compares the brittle compressive strength predicted by the above model to the experimental measurements from ice and a variety of rocks deformed over a range of confinement. The

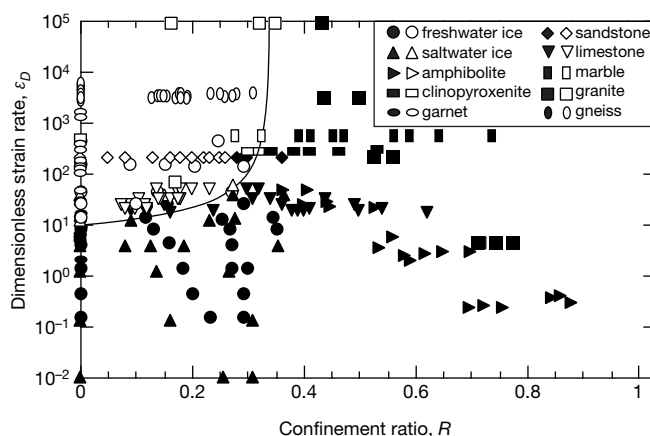


Figure 2 Deformation style as a function of dimensionless strain rate and confinement ratio. Black symbols indicate ductile behaviour, white symbols indicate brittle behaviour. The curve shows the predicted transition assuming $\mu = 0.5$. The temperatures at

which these data were collected are shown in Table 1. Data were taken from refs 15, 20–25, 28 and 30.

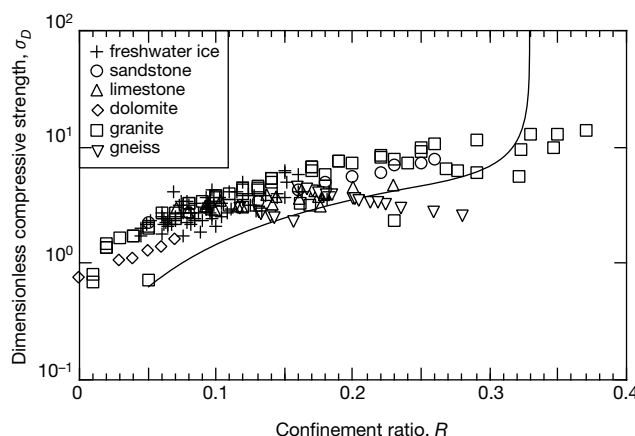


Figure 3 Brittle compressive strength of ice and various rock types as a function of confinement ratio. The curve shows the predicted compressive strength assuming

$\mu = 0.5$. The temperatures at which these data were collected are shown in Table 1. Data were taken from refs 14, 15 and 25–30.

secondary, or 'comb', crack model fits the data quite well over the regime of moderate confinement for which it was created, consistently matching the lower bound. Comparison of our model with previous models is difficult because of the use by other investigators of fitting parameters. For example, failure strengths predicted by the model of Horii and Nemat-Nasser¹⁹ are generally a factor of two to three greater than those predicted by our model. However, their model has three free parameters that describe the length and spacing distributions of the primary cracks at failure. Similarly, the models of Peng and Johnson⁴ and of Bazant and Xiang⁵ use the aspect ratio of microcolumns created by wing cracks emanating from adjacent primary cracks as a fitting parameter. None of these parameters is easily predicted a priori. If the microcolumn aspect ratio predicted by equations (3) and (4) (which is consistent with laboratory observations) is used in the models of Peng and Johnson⁴ and Bazant and Xiang⁵, which assume that microcolumn failure occurs via elastic buckling, the predicted compressive strengths are two to three orders of magnitude greater than measured values.

The agreement between prediction and measurement (Figs 2 and 3) provides support for the major role of secondary cracks in brittle compressive failure. The confinement ratio and strain rate required to initiate secondary cracks control the brittle-ductile transition, whereas the bending-induced failure of the microcolumns created between secondary cracks controls strength. In addition, the agreement between theory and experiment for a variety of materials suggests that micro-mechanical processes similar to those described here are active in a variety of crystalline solids. The new model, therefore, should have broad applicability. □

Received 20 December 2000; accepted 9 July 2001.

1. Brace, W. F. & Bombolakis, E. G. A note on brittle crack growth in compression. *J. Geophys. Res.* **68**, 3709–3713 (1963).
2. Horii, H. & Nemat-Nasser, S. Brittle failure in compression: Splitting, faulting, and brittle-ductile transition. *Phil. Trans. R. Soc. Lond. A* **319**, 337–374 (1986).
3. Ashby, M. F. & Hallam, S. D. The failure of brittle solids containing small cracks under compressive stress states. *Acta Metall.* **34**, 497–510 (1986).
4. Peng, S. & Johnson, A. M. Crack growth and faulting in cylindrical specimens of Chelmsford Granite. *Int. J. Rock Mech. Min. Sci. Geomech. Abstr.* **9**, 37–86 (1972).
5. Bazant, Z. & Xiang, Y. Size effect in compression fracture: Splitting crack band propagation. *J. Eng. Mech.* **123**, 162–172 (1997).
6. Schulson, E. M., Iliescu, D. & Renshaw, C. E. On the initiation of shear faults during brittle compressive failure: A new mechanism. *J. Geophys. Res.* **104**, 695–705 (1999).
7. Reches, Z. & Lockner, D. A. Nucleation and growth of faults in brittle rocks. *J. Geophys. Res.* **99**, 18159–18173 (1994).
8. Riedel, H. & Rice, J. R. in *Fracture Mechanics: Twelfth conference* 112–130 (American Society for Testing and Materials, 1980).
9. Schulson, E. M. The brittle compressive fracture of ice. *Acta Metall. Mater.* **38**, 1963–1976 (1990).
10. Hellan, K. *Introduction to Fracture Mechanics* 1–302 (McGraw-Hill, New York, 1984).
11. Qi, S. & Schulson, E. M. in *14th Int. Symp. on Ice* 521–527 (A. A. Balkema, Potsdam, New York, 1998).
12. Nixon, W. A. & Schulson, E. M. A micromechanical view of the fracture toughness of ice. *J. Physique* **48**, C1-313–C1-319 (1987).
13. Thouless, M. D., Evans, A. G., Ashby, M. F. & Hutchinson, J. W. The edge cracking and spalling of brittle plates. *Acta Metall.* **35**, 1333–1341 (1987).
14. Iliescu, D. *Contributions to Brittle Compressive Failure of Ice*. PhD Thesis, 1–284, Thayer School of Engineering, Dartmouth College (2000).
15. Heard, H. C. in *Rock Deformation, Mem. Geol. Soc. Am.* (eds Griggs, D. T. & Handin, J.) Vol. 79, 193–226 (Geol. Soc. Am., New York, 1960).
16. Weiss, J. & Schulson, E. M. Grain-boundary sliding and crack nucleation in ice. *Phil. Mag. A* **80**, 279–300 (2000).
17. Cooke, M. L. Fracture localization along faults with spatially varying friction. *J. Geophys. Res.* **102**, 22425–22434 (1997).
18. Pollard, D. D. & Segall, P. in *Fracture Mechanics of Rock* (ed. Atkinson, B. K.) 227–349 (Academic, San Diego, California, 1987).
19. Horii, H. & Nemat-Nasser, S. Compression-induced microcrack growth in brittle solids: axial splitting and shear fracture. *J. Geophys. Res.* **90**, 3105–3125 (1985).
20. Schulson, E. M. & Buck, S. E. The ductile-to-brittle transition and ductile failure envelopes of orthotropic ice under biaxial compression. *Acta Metall. Mater.* **43**, 3663–3668 (1995).
21. Schulson, E. M. & Nickolayev, O. Y. Failure of columnar saline ice under biaxial compression: Failure envelopes and the brittle-to-ductile transition. *J. Geophys. Res.* **100**, 22383–22400 (1995).
22. Kirby, S. H. Tectonic stresses in the lithosphere: Constraints provided by the experimental deformation of rocks. *J. Geophys. Res.* **85**, 6353–6363 (1980).
23. Hacker, B. R. & Christie, J. M. in *The Brittle–Ductile Transition in Rocks* (eds Duba, A. G., Durham, W. B., Handin, J. W. & Wang, H. F.) (American Geophysical Union, Washington DC, 1990).
24. Wang, Z. & Ji, S. Deformation of silicate garnets: Brittle-ductile transition and its geological implications. *Can. Mineral.* **37**, 525–541 (1999).
25. Gowd, T. N. & Rummel, F. Effect of confining pressure on the fracture behaviour of a porous rock. *Int. J. Rock Mech. Min. Sci. Geomech. Abstr.* **17**, 225–229 (1980).

26. Mogi, K. Some precise measurements of fracture strength of rocks under uniform compressive stress. *Rock Mech. Eng. Geol.* **4**, 41–55 (1966).
27. Shock, R. N. & Heard, H. C. Static mechanical properties and shock loading response of granite. *J. Geophys. Res.* **79**, 1662–1666 (1974).
28. Tullis, J. & Yund, R. A. Experimental deformation of dry Westerly granite. *J. Geophys. Res.* **82**, 5705–5718 (1977).
29. Brace, W. F. & Byerlee, J. D. in *Failure and Breakage of Rock* (ed. Fairhurst, C.) (American Institute of Mining, Metallurgical and Petroleum Engineers, New York, 1967).
30. Gottschalk, R. R., Kronenberg, A. K., Russell, J. E. & Handin, J. Mechanical anisotropy of gneiss: Failure criterion and textural sources of directional behavior. *J. Geophys. Res.* **95**, 21613–21634 (1990).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

This work was supported by the US National Science Foundation, the Army Research Office, and the Office of Naval Research.

Correspondence and requests for materials should be addressed to C.E.R. (e-mail: carl.renshaw@dartmouth.edu) or to E.M.S. (e-mail: erland.schulson@dartmouth.edu).

Transition of Mount Etna lavas from a mantle-plume to an island-arc magmatic source

Pierre Schiano*, Roberto Clocchiatti†, Luisa Ottolini‡ & Tiziana Busà§

* Laboratoire "Magmas et Volcans", Université Blaise Pascal, CNRS UMR 6524, OPGC, 5 rue Kessler, 63038 Clermont-Ferrand, France

† Laboratoire Pierre Süe, CEA CNRS UMR 9956, Centre d'Etude Nucléaire de Saclay, 91191 Gif-sur-Yvette, France

‡ Centro di Studio per la Cristallografia e la Cristallografia-CNR, via Ferrata 1, I-27100 Pavia, Italy

§ Dipartimento di Scienze Geologiche, Università di Catania, Corso Italia 55, I-95129 Catania, Italy

Mount Etna lies near the boundary between two regions that exhibit significantly different types of volcanism. To the north, volcanism in the Aeolian island arc is thought to be related to subduction of the Ionian lithosphere¹. On Sicily itself, however, no chemical^{2,3} or seismological⁴ evidence of subduction-related volcanism exists, and so it is thought that the volcanism—including that on Mount Etna itself—stems from the upwelling of mantle material⁵, associated with various surface tectonic processes^{1,6}. But the paucity of geological evidence regarding the primary composition of magma from Mount Etna means that its source characteristics remain controversial. Here we characterize the trace-element composition of a series of lavas emitted by Mount Etna over the past 500 kyr and preserved as melt inclusions inside olivine phenocrysts. We show that the compositional change in primary magmas from Mount Etna reflects a progressive transition from a predominantly mantle-plume source to one with a greater contribution from island-arc (subduction-related) basalts. We suggest that this is associated with southward migration of the Ionian slab, which is becoming juxtaposed with a mantle plume beneath Sicily. This implies that the volcanism of Mount Etna has become more calc-alkaline, and hence more explosive, during its evolution.

Volcanoes in eastern Sicily (Mount Etna and the volcano-sedimentary Iblean plateau) are situated on the suture between the converging European and African plates, adjacent to the subduction-related Aeolian island arc¹ (Fig. 1). There is, however, no clear evidence that this volcanism is directly connected with active subduction beneath the Aeolian arc; that is, there is no subduction-

related seismic trace beneath Sicily⁴. It has therefore been proposed that Sicily is not located directly above the active subduction zone but at the lateral edge¹. Volcanic rocks of the Iblean Mountains range from quartz-tholeiites to undersaturated alkaline lavas², and volcanism at Mount Etna is composed for the most part of intermediate alkaline lavas³. On the other hand, a wide range of rock types occurs within the nearby Aeolian islands, from typical island-arc calc-alkaline series to shoshonitic potassic lavas⁷. The coexistence of alkaline-type and calc-alkaline-type volcanism places a number of constraints on any geodynamic picture of the southern Tyrrhenian Sea. It has for example been proposed that the Aeolian calc-alkaline series is associated with subduction of Ionian lithosphere, whereas alkaline-type volcanism in eastern Sicily is related to extensional tectonics of a corresponding backarc marginal basin (see, for example, ref. 1). Alternatively, Mount Etna has been recently attributed to sideways suction of asthenospheric material from beneath the African plate, induced by rollback motion of the Ionian slab⁶.

The samples in which the primitive olivine-hosted melt inclusions were analysed come from Vulcano island (in the Aeolian island arc), the Iblean Mountains and Mount Etna (Fig. 1). Etnean samples represent different periods of volcanism. They include tholeiitic hyaloclastites from the Aci Castello series, which represents samples of the oldest Etnean volcanics so far identified (520 ± 40 kyr before present, BP)⁸, basalts from the oldest part of the Etnean alkaline sequence (Santa Venera deposits dated at ~ 120 kyr BP⁹), alkali basalts from Mount Maletto (the most primitive of all Etnean holocene lavas), which represent recent (~ 7 kyr BP)¹⁰ eccentric flank eruptions, and present-day (February 1999 eruption)¹¹ hawaiite scoria from the southeast crater on the summit of Mount Etna. Basanite samples collected in the breccia deposits from Villasmundo (Iblean Mountains) have been dated to about 7.38 Myr ago (ref. 12) and high-K basaltic samples from a scoria cone on La Sommata (Vulcano island) were erupted ~ 50 kyr ago (ref. 13). Except for the 1999 eruption of the southeast crater on

the summit of Mount Etna, the selected samples are the most primitive (in terms of MgO content) basalts and contain highly magnesian olivine crystals (85–91% forsterite; FO_{85-91}) as the main phenocryst type.

The major- and trace-element compositions of melt inclusions (after homogenization or correction for post-entrapment crystallization, see Methods) are given in Supplementary Information. The inclusions range in composition from nepheline-normative alkalic basalt to hypersthene-normative transitional basalt. They show a large range of concentrations for light rare-earth elements and high-field-strength elements (HFSE), suggesting a large range of parental magma compositions (Fig. 2). However, correlations between ratios involving highly incompatible elements (with bulk solid/melt partition coefficients, $D \ll$ melt fraction; Fig. 3) indicate that the trace-element variations among melt inclusions reflect a mixing relationship between two mantle sources, where one end-member is a mantle component represented by the Iblean Mountains and the other is a mantle component represented by Vulcano island.

We now consider these two mantle components (defined by the mixing arrays of Fig. 3) in more detail. Olivine-hosted melt inclusions from Vulcano have high concentrations of large-ion lithophile elements, high ratios of light-rare-earth elements to heavy-rare-earth elements, and relatively low concentrations of HFSE (Fig. 2). Such trace-element features are typical of arc basalts, and are thought to reflect melting of mantle sources that have been metasomatized by fluid and sediment components derived from the downgoing slab. In contrast, melt inclusions from the Iblean Mountains lack the negative HFSE anomalies considered to be evidence for a subduction-related origin. They require the existence of a mantle source enriched in incompatible elements relative to normal mid-ocean-ridge basalts (N-MORB). Their trace-element compositions are remarkably similar to the ocean island basalts (OIB) showing HIMU-type isotopic signatures (very high Pb and relatively low Sr isotope ratios; Figs 2 and 3); that is, they display a continuous decrease in the normalized abundances of Ba, K and Rb, and enrichments of Nb relative to similarly incompatible elements^{14,15}. The involvement of a HIMU-type mantle component in the sources of basalts from the Iblean Mountains and Mount Etna is also confirmed by recent Pb–Sr–Nd–Hf isotopic data¹⁶. HIMU-type basalts are generally associated with mantle plumes, and their trace-element and isotopic features are thought to be derived from a source that may incorporate large segments of subduction-zone-processed recycled oceanic crust (see, for example, ref. 15).

The characterization of a HIMU-type component, which is generally attributed to deep-rooted mantle plumes, in the source of early Etnean eruptions of Aci Castello gives additional constraints on the origin of Mount Etna. In particular, involvement of a plume-type mantle component cannot be reconciled with shallow-level extensional tectonics (see ref. 17 and references therein). In contrast, it supports the hypothesis that early Etnean emissions (and the volcano-sedimentary Iblean plateau) are caused by a mantle plume located in a globally convergent geodynamical setting (see, for example, ref. 5). Note, however, that this model explains how plume-type and arc-type mantle sources can co-exist at the scale of the southern Tyrrhenian–Sicilian region, but does not clarify their interaction (as illustrated by the mixing trends of Fig. 3).

Volcanic activity of Mount Etna is characterized by a time-stratigraphic progression from primitive lavas with tholeiitic affinities (for example, the submarine Aci Castello series) to more alkaline rocks (including basalts from Santa Venera and Mount Maletto)³. This could be a similar transition to that observed on Hawaiian volcanoes. However, the tholeiitic-alkalic transition in Hawaii is marked by a gradual decrease in the degree of melting of a relatively uniform source (see, for example, ref. 18). In contrast, the very different trace-element signatures for olivine-hosted melt inclusions in Aci Castello, Santa Venera and Mount Maletto basalts (Fig. 2) point towards a change of source material during the

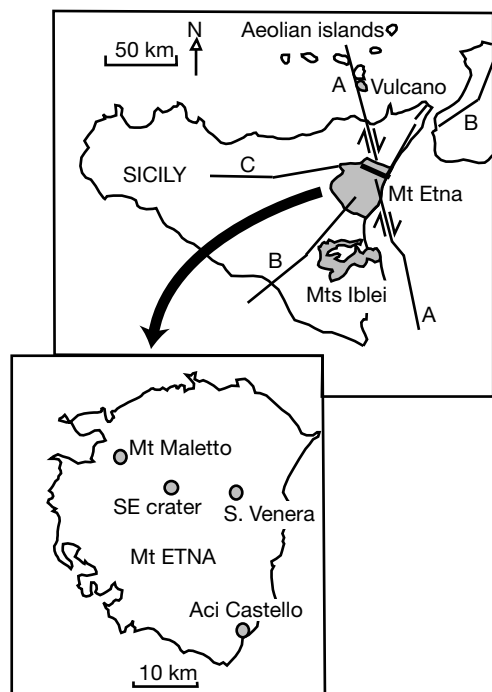


Figure 1 Sketch map of islands in the southern Tyrrhenian Sea and sample locations on Mount Etna. Solid lines indicate major strike-slip fault systems²⁵: A, Tindari–Giardini–Etna–Malta escarpment; B, Comiso–Etna–Messina–Capo Vaticano; C, Monte Kumeta–Alcantara.

magmatic evolution of Mount Etna. Melt inclusions from Aci Castello and Santa Venera display the trace-element characteristics of plume-derived HIMU basalts, whereas melt inclusions from Mount Maletto have trace-element compositions close to the primitive melt inclusions from subduction-related Vulcano island. Moreover, the mixing trends of Fig. 3 show that Mount Maletto melt inclusions represent liquids between typical OIB-type and island-arc-basalt type primitive magmas, represented respectively by melt inclusions from the Iblean Mountains and Vulcano island.

These observations provide strong evidence that the magma source beneath Mount Etna was originally plume-derived and has progressively become more arc-like. The melt inclusion data presented here therefore have implications for subduction-related magmatism—in particular, they are relevant to the debate about a possible contribution of mantle plumes to island-arc lavas (see, for example, refs 19, 20). Not only does the plume–arc transition process recorded by the Etnan primitive melt inclusions represent, to our knowledge, the first documentation of such a source evolution beneath a single volcano, but the respective signatures of Santa

Venera and Mount Maletto melt inclusions indicate that the evolution has occurred in less than 100 kyr. It should also be noted that, at first sight, trace-element data for Mount Maletto melt inclusions can be accounted for by melting an OIB-like mantle wedge, provided that a HFSE-rich phase is residual during partial melting²¹. But we consider this hypothesis unlikely, because Nb remains a highly incompatible element (that is, the bulk D for Nb is equivalent to that of Ce, see Fig. 2 legend) during the melting process recorded by the melt inclusions. Therefore, the low Nb concentrations in the Mount Maletto melt inclusions do not reflect the presence of HFSE-rich residual phases in an enriched mantle wedge, but rather the low abundance of Nb in the mantle sources of the melts.

The recent (post-1970) magmatic activity of Mount Etna is characterized by a higher frequency of eruption and higher effusion rates²². Moreover, the composition of the lavas shows anomalous enrichments in alkalis²³ with an accompanying increase of $^{87}\text{Sr}/^{86}\text{Sr}$ ratios (see, for example, ref. 10). Olivine (Fo_{81})-hosted melt inclusions in hawaiite scoria of the 4 February 1999 eruption from the

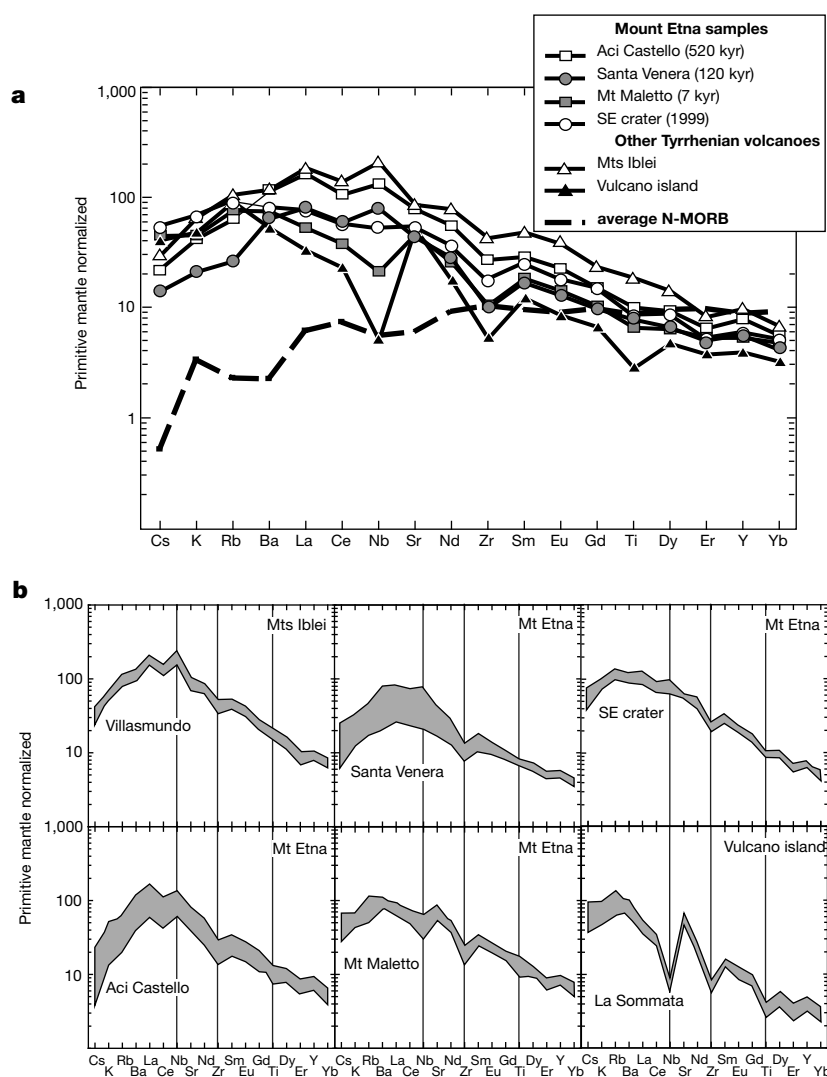


Figure 2 Trace-element diagrams. Olivine-hosted melt inclusions in Etnan basalts belonging to suites of different age are compared with olivine-hosted melt inclusions in basalts from other volcanoes of the southern Tyrrhenian–Sicilian region (Vulcano island and the Iblean Mountains). **a**, Representative trace-element patterns for melt inclusions from Mount Etna (Aci Castello, Santa Venera, Mount Maletto and the 1999 southeast crater series), Vulcano island and the Iblean Mountains. The order of element corresponds to the incompatibility sequence (taken as the order of their bulk solid/melt partition

coefficients), defined by considering the element concentrations in the magmatic series and using standard deviations of the element concentration averages²⁸. Compositions are normalized to the primitive mantle composition of ref. 29. Also shown for comparison is the average N-MORB (normal mid-ocean-ridge basalt) composition from ref. 29.

b, Ranges of trace-element compositions (normalized as in **a**) for olivine-hosted melt inclusions in basalts from Mount Etna (Aci Castello, Santa Venera, Mount Maletto and the 1999 southeast crater series), Vulcano island and the Iblean Mountains.

southeast crater¹¹ were then analysed for comparison. When plotted in the diagrams of Figs 2 and 3, the 1999 melt inclusions have strong similarities with Mount Maletto melt inclusions; that is, their trace-element patterns and highly incompatible element ratios are intermediate between Vulcano and Iblean Mountains melt inclusions. This suggests that the plume–arc interaction process recorded by the olivine-hosted primitive melt inclusions from Mount Maletto basalts can also account for the geochemical peculiarities of the present-day Etnean lavas. Therefore, the K and Rb enrichments and the gradual ⁸⁷Sr/⁸⁶Sr increase in Etnean historical lavas could simply reflect increasing participation of the arc-type source component. This hypothesis is consistent with the high H₂O and S concentrations of present-day Etnean magmas, which are more akin to subduction-related volcanoes than to oceanic hotspot volcanoes²⁴.

It is important to relate the geochemical evidence for a general transition from plume-type to arc-type mantle sources beneath Mount Etna to the regional setting. A simple model would be that subduction-related magmas originate from the nearby Aeolian island arc and migrate along tectonic features, such as the north-west–southeast Tindari–Giardini–Etna–Malta fault system (A–A in Fig. 1; see also ref. 25) to interact with Etnean plume-related magmas. This hypothesis, however, is weakened by the observation that, to our knowledge, Etnean lavas with plume signature have only been identified in the earlier succession. An alternative hypothesis is that the southwestern lateral limit of the Ionian slab beneath Mount Etna is progressively shifting. This could juxtapose subduction zone and mantle plume, allowing their respective melts to provide the mixed-source magmas. Finally, the plume–arc interaction process beneath Mount Etna may explain the characteristic features of its recent volcanic activity when compared with a typical hotspot

volcano. In particular, Mount Etna displays an unusual eruptive style, characterized by permanent strombolian explosions sometimes culminating in lava fountains (see, for example, ref. 26), and abundant gas emissions. Because subduction-related volcanoes are highly explosive, understanding the triggering mechanism of the plume–arc transition process identified here (and anticipating its evolution) is thus important when assessing the nature of future explosive events at Mount Etna. □

Methods

Melt inclusions

Because olivine-hosted melt inclusions in hyaloclastites (from Aci Castello) and pyroclastic scoria (from Vulcano island and the southeast crater on the summit of Mount Etna) were naturally quenched during eruption, post-entrapment reactions that occurred inside the inclusions are restricted to slight precipitation of host phases on inclusion walls. Correction for this post-entrapment crystallization was made by incrementally adding olivine in Fe/Mg equilibrium until the liquidus olivine is identical to the host olivine. In all cases the amount of host olivine added to the glass inclusions was less than 6% by weight. Note, however, that the extent of the correction is not critical for the results, as this procedure simply induces a dilution by a constant factor of all the incompatible trace elements discussed here—so without correction, the relative trace-element patterns of the melt inclusions would be unchanged. For olivine-hosted melt inclusions in effusive flows from Mount Maletto, Santa Venera, and the Iblean Mountains, cooling rates were slow enough to induce both precipitation of the host olivine at the inclusion walls and nucleation of additional minerals inside the inclusions. The original melt compositions were produced in the glass samples after heating each sample up to the temperature at which the inclusion content homogenizes to a uniform single melt phase.

Analytical techniques

Major-element compositions of homogenized and unheated glass inclusions were measured with a Cameca SX 50 electron microprobe at the Camparis Center (University of Paris VI), using an accelerating voltage of 15 keV, sample current of 10 nA and a defocused

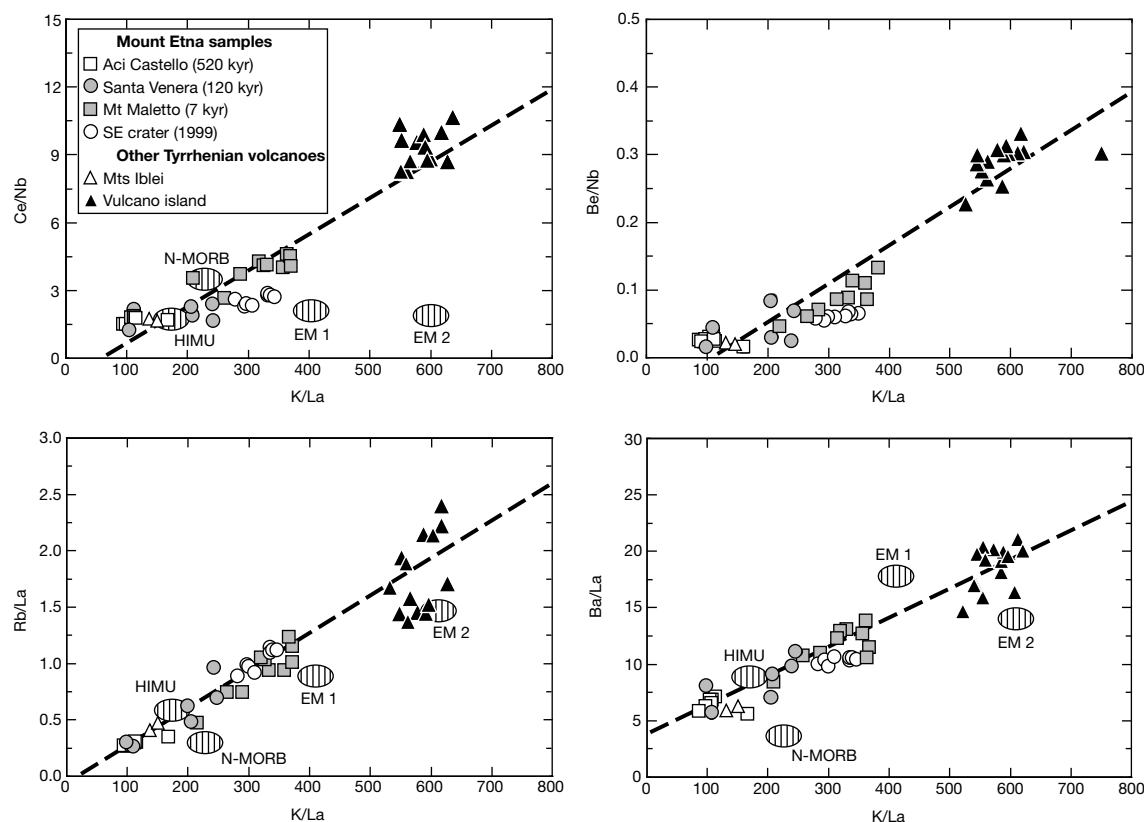


Figure 3 Plots of various ratios involving highly incompatible elements. Melt inclusions in olivine phenocrysts in dated basalts representative of the major episodes of Mount Etna volcanism are compared with melt inclusions in olivine phenocrysts in basalts from other volcanoes of the southern Tyrrhenian–Sicilian region. Also shown for comparison is

the estimated³⁰ composition of the sources of HIMU (see text), EM1 and EM2-type (the two other isotopically extreme signatures) OIB, and the estimated²⁹ average N-MORB composition.

beam, and trace-element compositions were measured with a Cameca IMS 4f ion microprobe at CNR-CSCC (Pavia, Italy), following the techniques described in ref. 27. Moreover, Sc and Cr were corrected for interferences by $^{29}\text{Si}^{16}\text{O}^+$ and $^{24}\text{Mg}^{28}\text{S}^+$, at mass 45 and 52 (a.m.u.), respectively.

Received 5 February; accepted 5 July 2001.

- Barberi, F., Gasparini, P., Innocenti, F. & Villari, L. Volcanism of the southern Tyrrhenian Sea and its geodynamic implications. *J. Geophys. Res.* **78**, 5221–5232 (1974).
- Bianchini, G., Clocchiatti, R., Coltorti, M., Joron, J. L. & Vaccaro, C. Petrogenesis of mafic lavas from the northernmost sector of the Iblean district (Sicily). *Eur. J. Mineral.* **10**, 301–315 (1998).
- Tanguy, J. C. Tholeiitic basalt magmatism of Mount Etna and its relations with the alkaline series. *Contrib. Mineral. Petrol.* **66**, 51–67 (1978).
- Anderson, H. & Jackson, J. The deep seismicity of the Tyrrhenian Sea. *Geophys. J. R. Astron. Soc.* **91**, 613–637 (1987).
- Tanguy, J. C., Condomines, M. & Kieffer, G. Evolution of the Mount Etna magma: Constraints on the present feeding system and eruptive mechanism. *J. Volcanol. Geotherm. Res.* **75**, 221–250 (1997).
- Gvirtzman, Z. & Nur, A. The formation of Mount Etna as the consequence of slab rollback. *Nature* **401**, 782–785 (1999).
- Keller, J. Petrology of some volcanic rock series of the Aeolian Arc, Southern Tyrrhenian Sea: Calc-alkaline and shoshonitic associations. *Contrib. Mineral. Petrol.* **45**, 29–47 (1974).
- Gillot, P. Y., Kieffer, G. & Romano, R. The evolution of Mount Etna in the light of potassium-argon dating. *Acta Vulcanol.* **5**, 81–87 (1994).
- Busà, T., Clocchiatti, R. & Cristofolini, R. Low P and high P alkaline magmas from Mt Etna: their origin and evolution from the study of mineralogy and chemistry of lavas and of olivine-trapped inclusions. *Periodico Mineralogia* **68**, 163–183 (1999).
- Condomines, M., Tanguy, J. C. & Michaud, V. Magma dynamics at Mt Etna: Constraints from U-Th-Ra-Pb radioactive disequilibria and Sr isotopes in historical lavas. *Earth Planet. Sci. Lett.* **132**, 25–41 (1995).
- La Delfa, S., Patané, G., Clocchiatti, R., Joron, J. L. & Tanguy, J. C. Activity of Mount Etna preceding the February 1999 fissure eruption: inferred mechanism from seismological and geochemical data. *J. Volcanol. Geotherm. Res.* **105**, 121–139 (2001).
- Trua, T. *Evoluzione del magmatismo ibleo (Sicilia sud orientale) dal Cretaceo al Plio-Pleistocene*. Thesis, Univ. Pisa (1997).
- Gioncada, A. et al. A study of melt inclusions at Vulcano (Aeolian Islands, Italy): insights on the primitive magmas and on the volcanic feeding system. *Bull. Volcanol.* **60**, 286–306 (1998).
- Weaver, B. L. The origin of ocean-island basalt end-member compositions: trace element and isotopic constraints. *Earth Planet. Sci. Lett.* **104**, 381–397 (1991).
- Hofmann, A. W. Mantle geochemistry: the message from oceanic volcanism. *Nature* **385**, 219–229 (1997).
- Albarede, F. et al. OIB-type magmas in subduction zones: mantle counterflow above detaching plates. *Eos* **81**, 1271 (2000).
- Ferrari, L. & Manetti, P. Geodynamic framework of the Tyrrhenian volcanism: a review. *Acta Vulcanol.* **3**, 1–19 (1993).
- Feigenson, M. D., Hofmann, A. W. & Spera, F. J. Case studies on the origin of basalt. II. The transition from tholeiitic to alkalic volcanism on Kohala volcano, Hawaii. *Contrib. Mineral. Petrol.* **84**, 390–405 (1983).
- Wendt, J. L., Regelous, M., Collerson, K. D. & Ewart, A. Evidence for a contribution from two mantle plumes to island-arc lavas from northern Tonga. *Geology* **25**, 611–614 (1997).
- Turner, S. & Hawkesworth, C. Using geochemistry to map mantle flow beneath the Lau Basin. *Geology* **26**, 1019–1022 (1998).
- Morris, J. D. & Hart, S. R. Isotopic and incompatible element constraints on the genesis of island arc volcanics from Cold Bay and Amak Island, Aleutians, and implications for mantle structure. *Geochim. Cosmochim. Acta* **47**, 2015–2030 (1983).
- Tanguy, J. C., Kieffer, G. & Patané, G. Dynamics, lava volume and effusion rate during the 1991–1993 eruption of Mount Etna. *J. Volcanol. Geotherm. Res.* **71**, 259–265 (1996).
- Joron, J. L. & Treuil, M. Etude géochimique et pétrogénèse de laves de l'Etna, Sicile, Italie. *Bull. Volcanol.* **47**, 1125–1144 (1984).
- Metrich, N. & Clocchiatti, R. Sulfur abundance and its speciation in oxidized alkaline melts. *Geochim. Cosmochim. Acta* **60**, 4151–4160 (1996).
- Lanzafame, G. & Bousquet, J. C. The Maltese escarpment and its extension from Mt. Etna to the Aeolian Islands (Sicily): importance and evolution of a lithosphere discontinuity. *Acta Vulcanol.* **9**, 113–120 (1997).
- Cottelli, M., Del Carlo, P. & Vezzoli, L. Discovery of a Plinian basaltic eruption of Roman age at Etna Volcano, Italy. *Geology* **26**, 1095–1098 (1998).
- Bottazzi, P., Ottoloni, L., Vannucci, R. & Zanetti, A. in *Secondary Ion Mass Spectrometry SIMS IX* (eds Benninghoven, A., Niki, Y., Shimizu, R. & Werner, H. W.) 927–930 (Wiley & Sons, Chichester, 1994).
- Schiano, P., Allègre, C. J., Dupré, B., Lewin, E. & Joron, J. L. Variability of trace elements in basaltic suites. *Earth Planet. Sci. Lett.* **119**, 37–51 (1993).
- Hofmann, A. W. Chemical differentiation of the Earth: the relationship between mantle, continental crust, and oceanic crust. *Earth Planet. Sci. Lett.* **90**, 297–314 (1988).
- Allègre, C. J., Schiano, P. & Lewin, E. Differences between oceanic basalts by multitrace element ratio topology. *Earth Planet. Sci. Lett.* **129**, 1–12 (1995).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank J. C. Tanguy, A. Borgia, O. Sigmarsson, R. Cristofolini and B. Van Wyk de Vries for discussions, A. Kent for comments on the manuscript, and the Consiglio Nazionale delle Ricerche (Italy) for funding the ion microprobe that we used at CSCC (Pavia).

Correspondence and requests for materials should be addressed to P.S. (e-mail: schiano@opgc.univ-bpclermont.fr).

Genetic linkage of ecological specialization and reproductive isolation in pea aphids

David J. Hawthorne* & Sara Via*†

* Department of Entomology; and † Department of Biology, University of Maryland, College Park, Maryland 20742, USA

The evolution of ecological specialization generates biological diversity and may lead to speciation^{1–3}. Genetic architecture can either speed or retard this process. If resource use and mate choice have a common genetic basis through pleiotropy or close linkage, the resulting genetic correlations can promote the joint evolution of specialization and reproductive isolation, facilitating speciation^{4–6}. Here we present a model of the role of genetic correlations in specialization and speciation, and test it by analysing the genetic architecture of key traits in two highly specialized host races of the pea aphid (*Acyrtosiphon pisum pisum*; Hemiptera: Aphididae). We found several complexes of pleiotropic or closely linked quantitative trait loci (QTL) that affect key traits in ways that would promote speciation: QTL with antagonistic effects on performance on the two hosts are linked to QTL that produce assortative mating (through habitat choice). This type of genetic architecture may be common in taxa that have speciated under divergent natural selection.

Natural selection is increasingly recognized as the main cause of speciation, yet the detailed ecological and genetic mechanisms of the process remain obscure^{2,7}. To understand how adaptation to different environments drives speciation we need to re-examine the extent and mechanism of genetic correlations among characters within and across environments¹⁵. Genetic correlations (owing to pleiotropy or linkage disequilibrium) can either facilitate or constrain adaptive evolution⁸. For example, theory suggests that genetic correlations between resource use and mate choice can facilitate both specialization and speciation under divergent selection^{4–6}. However, measures of genetic correlations among key traits have not been much used in previous empirical studies of the genetics of speciation, which have focused primarily on mechanisms of hybrid sterility and inviability⁹. Consequently, we lack empirical evidence of the genetic architectures that theory has identified as favourable for speciation.

Two classes of genetic correlations are particularly important for speciation under divergent selection. First, fitness trade-offs (negative genetic correlations between traits that are under divergent selection in different environments) suggest one mechanism for ecological specialization^{3,10,11,12}. Secondly, positive genetic correlations between such fitness traits within environments and the characters that affect mate choice (including choice of mating habitat) can produce assortative mating among individuals specialized to the same environment, promoting reproductive isolation^{4–6,13,14}.

We integrated the genetic correlations of traits affecting specialization and mate choice into a single network. This revealed the potential for a set of complementary correlated responses to divergent selection that could facilitate specialization and reproductive isolation, even if the values of the individual genetic correlations are modest (Fig. 1a). Because every individual carries genes that affect traits in both environments, selection on resource use within environments produces correlated responses in traits that would be expressed in the alternative environment¹⁵. These correlated responses combine to increase the divergence and isolation of populations. Further direct and correlated responses are added if divergent selection also acts directly on the traits that lead

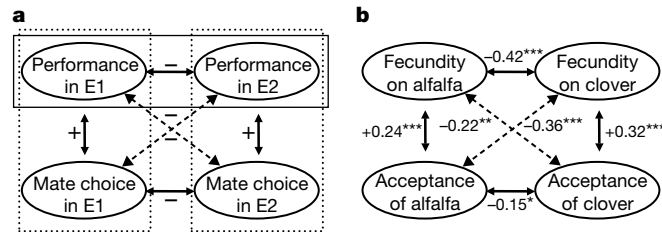


Figure 1 Network of complementary genetic correlations affecting specialization and speciation. **a**, Graphical model of how genetic correlations (negative/positive signs indicated next to arrows) can facilitate ecological specialization and reproductive isolation. A negative genetic correlation between performance in two environments (solid box) typifies a fundamental genetic 'trade-off' that facilitates the evolution of specialization. Positive genetic correlations between resource use and mate choice (dotted boxes) speed up the evolution of specialization and cause reproductive isolation between divergent populations. Correlations between performance and mate choice across environments

(dashed lines) have not been explicitly considered in previous models, but would add to the network of complementary correlated responses. **b**, Estimates of genetic correlations among key traits within a segregating F_2 progeny of a cross between pea aphids specialized to alfalfa and red clover. Fecundity measures the degree of ecological specialization, while differences in acceptance of each plant lead to mate choice and assortative mating. All traits are selected to increase¹⁸, leading to a network of complementary correlated responses that would facilitate population divergence. Triple asterisk, $P < 0.01$; double asterisk, $P < 0.05$; single asterisk, $P < 0.08$.

to habitat or mate choice. A genetic architecture that leads to this pattern of complementary genetic correlations among key traits could lead to the rapid reproductive isolation of specialists selected in different environments, regardless of whether divergence takes place in sympatry or allopatry.

To test this model, we measured the extent and mechanisms of the genetic correlations among resource use and mate choice in a cross between two incipient species of pea aphids that are highly specialized on alfalfa and red clover (both are in the subspecies *Acyrtosiphon pisum pisum*, Hemiptera: Aphididae¹⁶). These incipient species are found sympatrically and they exhibit genetically based trade-offs in performance on the two hosts^{11,17}. They are reproductively isolated primarily by differences in habitat acceptance, which leads to assortative mating because pea aphids mate exclusively on their host. Strong divergent individual selection acts both on resource use and on habitat acceptance^{11,18}, and so a genetic architecture causing a complementary pattern of correlations in the key traits (Fig. 1a) could have led to extremely rapid divergence of pea aphids selected on alfalfa and clover.

We produced F_2 hybrids from a cross between a pair of typical, highly specialized genotypes from the two incipient species, and measured the four key traits depicted in Fig. 1a in replicate for each F_2 genotype. We used early fecundity when reared on alfalfa (FecA) or on clover (FecC) as an indicator of specialized resource use. We used the behavioural acceptance of alfalfa (AccA) or clover (AccC) of each F_2 in a no-choice situation as a proxy for mate choice. Because specialized individuals preferentially accept the habitat in which their performance is highest¹⁸, this leads to assortative mating.

The pattern of genetic correlations observed in these traits among the F_2 hybrids (Fig. 1b) is absolutely consistent with predictions from the model in Fig. 1a: the genetic correlation between performance in the two environments is negative, whereas the genetic correlation between mate choice (host acceptance) and fecundity on the host is positive for both environments. Significant negative genetic correlations were also observed between acceptance of one host and fecundity on the other (Fig. 1b). Although these correlations represent segregating (co)variance between just two parental genotypes, most of the variance in this system is between rather than within populations^{11,17}. Thus, the differences between these two genotypes are likely to capture most of the (co)variance among genes that cause the differences between the incipient species.

Depending on the underlying genetic architecture, these correlations could be either the cause or a consequence of the specialization and divergence of these incipient species. Genetic correlations due to linkage disequilibrium of closely linked genes or pleiotropy can be built up with little or no selection (respectively)^{19,20} and may

dramatically affect both the dynamics and trajectory of evolution^{8,19,20}. We term these 'fundamental' genetic correlations. In contrast, genetic correlations that are due to linkage disequilibrium among unlinked loci can only be maintained under very strong selection or in a subdivided population structure^{19,20}, and should therefore be considered an outcome rather than a cause of the evolution of specialization ('derived' genetic correlations).

To probe the effects of individual chromosome segments on ecological specialization and assortative mating, and to distinguish the causes of the observed genetic correlations in Fig. 1b, we used quantitative trait loci (QTL) analysis. We examined QTL for ecological specialization (fecundity on the two hosts) and mate choice (host acceptance). We asked whether these QTL are in close enough proximity on the linkage map and of the appropriate directionality to suggest that they could contribute to fundamental genetic correlations.

The pea aphid linkage map (Fig. 2) contains four linkage groups, which is consistent with cytological observations of four chromosomes in pea aphids²¹. We found QTL on all four linkage groups (Fig. 2). For these four traits, the identified QTL together explain only 10–57% of the genetic variance, which suggests that other QTL of small effect are also involved in both ecological specialization and habitat acceptance in pea aphids.

On linkage groups I and II, we found four complexes of QTL that affect several of the four key traits and are close enough together to indicate either pleiotropy or linkage. Two of these QTL complexes (W and Y, Fig. 2) are found only on the chromosome from the alfalfa parent, whereas the other two complexes (X and Z) involve chromosomes from both parents. We identify QTL in complexes X and Z as potentially homologous in the two genomes on the basis of their positions relative to the seven codominant markers used to align the two maps ('Codom XXX' on Fig. 2).

The QTL in these complexes all have positive effects on characters expressed on the 'native host', that is, QTL on IIa (the alfalfa homologue for linkage group II) increase fecundity/acceptance of alfalfa (Fig. 2); or they have negative effects on characters expressed on the alternate host, that is, the QTL on IIa decrease fecundity on clover. Moreover, the directionality of effects of the closely adjacent QTL are completely consistent with the pattern of correlations shown in Fig. 1. Within each genome, QTL have antagonistic effects on fecundity in the two environments or on fecundity on one host and acceptance of the other. In contrast, the adjacent QTL affect fecundity and acceptance of the same host in the same direction. The probability of such coordinated directionality occurring by chance alone is $P < 0.00006$ (ref. 22).

Previous work suggests a functional link between host acceptance and fecundity on the host²³. Thus, it is possible that each QTL

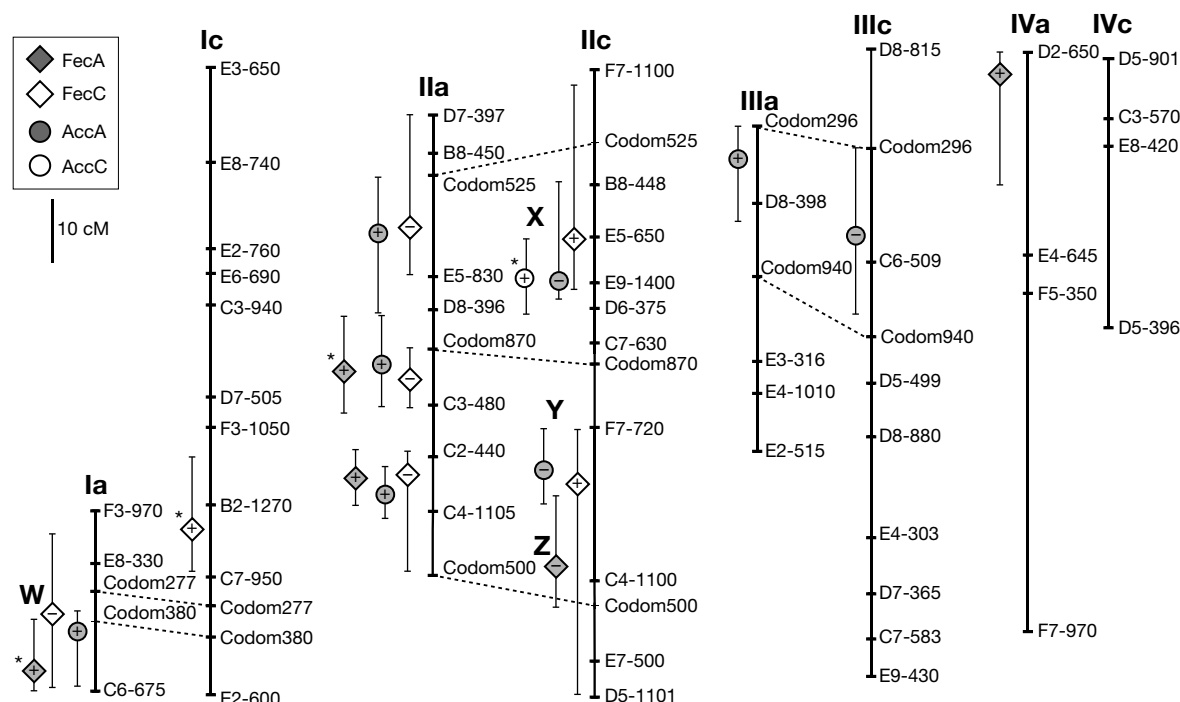


Figure 2 Linkage map of pea aphid, with quantitative trait loci (QTL) for ecological specialization and reproductive isolation. Linkage groups for each parental race are shown separately and indicated by a or c next to the linkage group number (for example, Ia is the first linkage group from the alfalfa genome). Linkage groups I–III were aligned using seven codominant markers as indicated by the dotted lines. FecA or FecC indicates fecundity in the first 9 days of adulthood on alfalfa or clover respectively. AccA or AccC

indicates behavioural acceptance of alfalfa or clover, respectively. QTL for the four key traits were placed on the map based on experimentwise significance levels of $P < 0.05$ from permutation tests: symbols marked with an asterisk suggest $0.08 < P < 0.12$; 95% confidence intervals for the location of each QTL were based on the permutation tests. The plus or minus sign within each symbol indicates the directionality of the QTL effect. Letters W, X, Y and Z mark complexes of QTL that affect several traits each.

complex represents a single gene with pleiotropic effects, although a higher-resolution analysis is required to verify this. However, even if the QTL are separate genes or gene groups, they are clearly closely linked (Fig. 2), and, for several reasons, they may function in genetic correlations almost as if they were pleiotropic. First, physical linkage greatly increases the probability that linkage disequilibrium will arise by genetic drift and decrease the eroding effects of recombination^{19,20}. For example, although linkage disequilibrium between unlinked loci declines by 50% in one generation of random mating, linkage disequilibrium between two loci that are 10 centimorgan (cM) apart would decline by only 9% (from equations (5.13) and (14.4) in ref. 20). Also, inbreeding, and especially cyclical parthenogenesis, further reduces the effects of recombination, with the effective map distance reduced from that under random mating in approximate inverse proportionality to the number of generations between recombination episodes¹⁹. As there are approximately ten clonal generations between each sexual generation of pea aphids, two loci that are 10–20 cM apart on the map would potentially recombine as if they were only 1–2 cM apart, greatly increasing the likelihood that they could enter into a fundamental genetic correlation. Finally, once a genetic correlation through linkage disequilibrium arises between host acceptance and fecundity on a host, perhaps by chance, it would provide an individual selective advantage and could be easily maintained and spread, catalysing further divergence.

Several other QTL were found that affect only one of the key traits. For example, a QTL on linkage group IVa increases fecundity on alfalfa but does not affect performance on clover. By contributing genetic variation to only one trait, such QTL reduce the overall genetic correlations induced by the pleiotropic/linked QTL, which may explain the modest correlations seen in Fig. 1b.

Our analyses provide evidence that the genetic architecture of

host use could have facilitated divergence and reproductive isolation of these incipient species. Although ecological specialization can occur in several different ways^{10,24}, none are as parsimonious as the complementary network of genetic correlations between resource use and mate choice (habitat acceptance) that is demonstrated here. Because this type of genetic architecture could significantly increase the probability of speciation for populations that evolve to utilize different resources, it may be a hallmark of ecological speciation by natural selection, especially in sympatric populations^{2,13,25}. □

Methods

Crosses

Pea aphids are cyclically parthenogenetic, so parental and F_2 genotypes can be clonally propagated. The mapping population was generated from pairwise reciprocal crosses between a single pair of clones, one a typical alfalfa specialist and the other a typical clover specialist. Two F_1 genotypes from this cross were reciprocally crossed to obtain 194 F_2 progeny. Each of these sexually produced genotypes was then maintained clonally, which permitted replicated phenotypic testing and increased the power of our QTL analysis.

Linkage mapping

Each F_2 was genotyped for 173 informative dominant AFLP markers, at which parents were determined to be homozygous for different alleles: all F_1 have a band (that is, they are heterozygous), and F_2 are 3:1 banded: unbanded. Separate linkage maps were made from recessive markers for each parental genotype using MapMaker/Exp²⁶ software. The two parental framework maps, made with a subset of the markers, were aligned for three of the four linkage groups using sequence-tagged AFLP fragments²⁷.

Demographic performance of F_2

The 194 F_2 progeny were tested in three or four replicates on each of the two hosts. For each trial, a winged adult of a given genotype was allowed to produce offspring for 24 h in an enclosure containing one of the hosts. The adult was then removed, and after 5 days, one of the offspring was moved to a new enclosure in which its fecundity in the first 9 days of adult life was measured (variables FecA or FecC, depending on the test host).

Behavioural acceptance of alfalfa and clover by F₂

For a given trial, four winged forms (alates¹⁷) of one F₂ genotype were placed onto a small enclosure containing one potted alfalfa or clover plant. Alates were all in their peak migratory period¹⁷. After 70 h, the percentage of alates in the enclosure that were on the host with offspring was used as the measure of host acceptance (AccA or AccC, for alfalfa and clover, respectively; all values were arcsine square-root transformed).

QTL mapping

Mapping analyses were performed on the best linear unbiased predictors (BLUPs, Proc Mixed, SAS²⁸) for each trait from the replicated trials: 9-day fecundity on each host (FecA, FecC), and behavioural acceptance of each host (AccA, AccC). We used composite interval mapping in QTL Cartographer²⁹ software, choosing markers as covariates for the analysis by stepwise regression. QTL for each character were placed on the linkage map if they were significant experimentwise at $P < 0.05$ by permutation test³⁰. A few putative QTL with lower significance levels ($0.08 < P < 0.12$) are also plotted (clearly marked on the figures by asterisks).

Received 15 February; accepted 5 July 2001.

1. Thompson, J. N. *The Coevolutionary Process* (Univ. Chicago Press, Chicago, 1994).
2. Schluter, D. in *Endless Forms: Species and Speciation* (eds Howard, D. J. & Berlocher, S. H.) 114–129 (Oxford Univ. Press, New York, 1998).
3. Futuyma, D. J. & Moreno, G. The evolution of ecological specialization. *Annu. Rev. Ecol. Syst.* **19**, 201–233 (1988).
4. Felsenstein, J. Skepticism toward Santa Rosalia, or why are there so few kinds of animals? *Evolution* **35**, 124–138 (1981).
5. Kondrashov, A. S. & Mina, M. V. Sympatric speciation, when is it possible? *Biol. J. Linn. Soc.* **27**, 201–233 (1986).
6. Rice, W. R. Selection via habitat specialization, the evolution of reproductive isolation as a correlated character. *Evol. Ecol.* **1**, 301–314 (1987).
7. Schemske, D. W. Understanding the origin of species. *Evolution* **54**, 1069–1073 (2000).
8. Lande, R. Quantitative genetic analysis of multivariate evolution, applied to brain: body size allometry. *Evolution* **33**, 402–416 (1979).
9. Orr, H. A. The population genetics of speciation: the evolution of hybrid incompatibilities. *Genetics* **139**, 1805–1813 (1995).
10. Fry, J. D. The evolution of host specialization, are trade-offs overrated? *Am. Nat.* **148**, S84–S107 (1996).
11. Via, S. The genetic structure of host plant adaptation in a spatial patchwork: demographic variability among reciprocally transplanted pea aphid clones. *Evolution* **45**, 827–852 (1991).
12. Filchak, K. E., Roethele, J. B. & Feder, J. L. Natural selection and sympatric divergence in the apple maggot *Rhagoletis pomonella*. *Nature* **407**, 739–742 (2000).
13. Bush, G. L. Sympatric speciation in animals: new wine in old bottles. *Trends Ecol. Evol.* **9**, 285–288 (1994).
14. Rice, W. R. & Hostert, E. Laboratory experiments on speciation. What have we learned in 40 years? *Evolution* **47**, 1637–1653 (1993).
15. Via, S. & Lande, R. Genotype–environment interaction and the evolution of phenotypic plasticity. *Evolution* **39**, 505–522 (1985).
16. Eastop, V. F. Keys for the identification of *Acyrtosiphon* (Hemiptera: Aphididae). *Bull. Br. Mus. Nat. Hist. B* **26**, 1–115 (1973).
17. Via, S. Reproductive isolation between sympatric races of pea aphids. I. Gene flow restriction and habitat choice. *Evolution* **53**, 1446–1457 (1999).
18. Via, S., Bouck, A. C. & Skillman, S. Reproductive isolation between sympatric races of pea aphids. II. Selection against migrants and hybrids in the parental environments. *Evolution* **54**, 1626–1637 (2000).
19. Lande, R. The genetic correlation between characters maintained by selection, linkage and inbreeding. *Genet. Res. Cambridge* **44**, 309–320 (1984).
20. Lynch, M. & Walsh, J. B. *Genetics and Analysis of Quantitative Traits* (Sinauer, Sunderland, Massachusetts, 1998).
21. Sun, R. Y. & Robinson, A. G. Chromosome studies of 50 species of aphids. *Can. J. Zool.* **44**, 649–653 (1966).
22. Orr, H. A. Testing natural selection vs. genetic drift in phenotypic evolution using quantitative trait locus data. *Genetics* **149**, 2099–2104 (1998).
23. Caillaud, C. M. & Via, S. Specialized feeding behavior influences both ecological specialization and assortative mating in sympatric host races of pea aphids. *Am. Nat.* **156**, 609–621 (2000).
24. Whitlock, M. C. The Red Queen beats the jack-of-all-trades? The limitations on the evolution of phenotypic plasticity and niche breadth. *Am. Nat.* **148**, S65–S77 (1996).
25. Via, S. Sympatric speciation in animals: the ugly duckling grows up. *Trends Ecol. Evol.* **16**, 381–390 (2001).
26. Lander, E. S. *et al.* MAPMAKER, an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* **1**, 174–181 (1987).
27. Cho, Y. G. *et al.* Cloning and mapping of variety specific rice genomic DNA sequences: Amplified fragment length polymorphisms (AFLP) from silver stained polyacrylamide gels. *Genome* **39**, 373–378 (1996).
28. Littell, R. C. *et al.* *SAS System for Mixed Models* (SAS Institute, Cary, North Carolina, 1996).
29. Basten, C. J., Weir, B. S. & Zeng, Z.-B. QTL Cartographer (Version 1.13) (Department of Statistics, North Carolina State University, Raleigh, North Carolina, 1996).
30. Churchill, G. & Doerge, R. W. Empirical threshold values for quantitative trait locus mapping. *Genetics* **138**, 963–971 (1994).

Acknowledgements

We thank G. Churchill, D. Schemske, D. Schluter and Z.-B. Zeng for discussions. M. Arnold, T. Bradshaw, J. Conner, K. Shaw and J. Wilkinson provided useful comments on the manuscript. Z.-B. Zeng provided the calculation from ref. 22, and R. Lande pointed

out the effects of cyclical parthenogenesis on recombination rates. A. Bouck, S. Skillman, C. Olson and A. Badgley provided expert technical assistance. This work was supported by NSF and USDA grants to S.V. and D.J.H.

Correspondence and requests for materials should be addressed to S.V. (e-mail: sv47@umail.umd.edu).

Dynamics of travelling waves in visual perception

Hugh R. Wilson*, Randolph Blake† & Sang-Hun Lee†

* *Biology and Centre for Vision Research, York University, 4700 Keele Street, Toronto, Ontario M3J 1P3, Canada*

† *Department of Psychology, Vanderbilt University, Nashville, Tennessee 37203, USA*

Nonlinear wave propagation is ubiquitous in nature, appearing in chemical reaction kinetics¹, cardiac tissue dynamics^{1,2}, cortical spreading depression³ and slow wave sleep⁴. The application of dynamical modelling has provided valuable insights into the mechanisms underlying such nonlinear wave phenomena in several domains^{1,2,5,6}. Wave propagation can also be perceived as sweeping waves of visibility that occur when the two eyes view radically different stimuli. Termed binocular rivalry, these fluctuating states of perceptual dominance and suppression are thought to provide a window into the neural dynamics that underlie conscious visual awareness^{7,8}. Here we introduce a technique to measure the speed of rivalry dominance waves propagating around a large, essentially one-dimensional annulus. When mapped onto visual cortex, propagation speed is independent of eccentricity. Propagation speed doubles when waves travel along continuous contours, thus demonstrating effects of collinear facilitation. A neural model with reciprocal inhibition between two layers of units provides a quantitative explanation of dominance wave propagation in terms of disinhibition. Dominance waves provide a new tool for investigating fundamental cortical dynamics.

On first experiencing binocular rivalry, people often comment not only on the remarkable disappearance of one monocular stimulus for several seconds at a time, but also on the highly ordered transitions in dominance as one stimulus sweeps the other out of conscious awareness. These dominance waves are particularly prominent with larger rival patterns subtending many degrees of visual angle⁹. To study these dominance waves we simplified pattern geometry by using annular stimuli, thereby restricting wave propagation effectively to the one dimension around the annulus. Readers capable of free-fusion may experience salient traveling waves using the rival patterns in Fig. 1a or the anaglyphs on the website (see Fig. 1 legend).

To measure the dynamics of wave propagation, we devised a technique allowing us to control the location at which dominance waves originate and, subsequently, to estimate travelling speed around the annulus. Our technique capitalizes on the fact that an abrupt contrast increment in a suppressed pattern reliably triggers its immediate dominance^{10,11}. The observer depressed and held the spacebar of a keyboard when the low-contrast radial grating was completely suppressed by the high-contrast spiral. This action triggered a brief contrast increment in the suppressed pattern at one of eight points (cardinal axes and diagonals). The spacebar was released only when the suppressed pattern became dominant at a monitored location, marked by nonius lines (see Methods). Data were consistent across all observers and demonstrated a linear

increase in propagation time with distance around the annulus (Fig. 1b). At the greatest distance around the annulus, two observers showed a flattening of their radial data, which is attributable to spontaneous reappearance of the suppressed pattern before arrival of the triggered dominance wave. For each observer propagation times, $T_p(x)$, were therefore fit with an equation incorporating constant-speed wave propagation (v) along with the gamma probability, $P(t)$, of spontaneous release from suppression at the target site before wave arrival:

$$T_p(x) = T_0 + \frac{x}{v} \left(1 - \int_0^{x/v} P(t) dt \right) + \int_0^{x/v} t P(t) dt \left(\int_0^{x/v} P(t) dt \right) \quad (1)$$

where x is the travel distance and T_0 is a constant response latency evident in the zero-distance data. The second and third terms in the equation are: (wave arrival time, x/v) $\times$ (probability of no prior spontaneous reappearance) + (expected time of spontaneous reappearance) $\times$ (probability of prior spontaneous reappearance). Least-mean-squares fitting of v and $P(t)$ parameters to the radial data revealed an average propagation speed across observers of 3.65 ± 0.54 degrees s^{-1} .

Recurrent excitatory connections in primate visual cortex preferentially interconnect cells with similar preferred orientations and

receptive fields that are roughly collinear^{12,13}. Both psychophysical¹⁴ and transcranial magnetic stimulation¹⁵ studies provide supporting evidence for collinear facilitation in humans. Therefore, we repeated our measurements using a low-contrast concentric target in place of the radial target. The same spiral pattern was again used, as it had the same local orientation difference (45°) from both radial and concentric contours. All observers again showed a linear increase in propagation time with distance, but the slopes were much shallower, signifying a greater propagation speed (Fig. 1b). Collinearity of the suppressed target contours increased speed approximately twofold in three observers and even more in the fourth, averaging 9.60 ± 4.76 degrees s^{-1} . This increase in speed is consistent with previous evidence for facilitation between collinear gratings during the dominance phase of rivalry¹⁶.

We investigated three additional aspects of dominance wave propagation. First, a low-contrast, spiral target pattern was used with a pitch angle orthogonal to the high-contrast spiral mask. This manipulation produced a speed of 5.8 degrees s^{-1} , intermediate between radial and concentric patterns (Fig. 1b, bottom left). Second, we tested whether dominance waves could propagate across a gap in the suppressed stimulus. Accordingly, a permanent gap (0.92° wide) in visual angle (three grating cycles) was introduced into the radial annulus at a point that was 67.5° distant from the marked arrival point. Dominance waves that were triggered 67.5° beyond the gap (that is, 135° from the arrival point) were blocked by the gap for both of the observers tested, and propagation times rose from 1.60 ± 0.055 s (S.L.) or 1.55 ± 0.056 s (R.B.) without the gap, to 2.71 ± 0.17 s (S.L.) or 2.28 ± 0.10 s (R.B.) in the presence of the gap. Both of these differences were highly significant ($t_{126} > 45.0$; $P < 10^{-6}$ for each subject), and the increased times correlate with the longer pathway (by 67%) in the opposite direction around the annulus when the shorter pathway is blocked by the gap. Very small gaps, however, can be traversed by dominance waves: a gap width of only 0.31° (one radial grating cycle) yielded equivalent propagation times for gap and no-gap conditions. Third, we determined whether eye movements would disrupt wave propagation. At the moment of wave initiation, observers shifted fixation from the central region (bull's-eye) of the target to the marked arrival point itself. Arrival times were now independent of distance around the annulus, implying that eye movements effectively abolished retinotopically-based wave propagation.

To learn how wave speed varies with eccentricity, we scaled our entire stimulus so that the mean annular radius doubled from 1.8 to 3.6° , spatial frequency being halved to compensate for reduced resolution at the greater eccentricity. Complete data using the radial

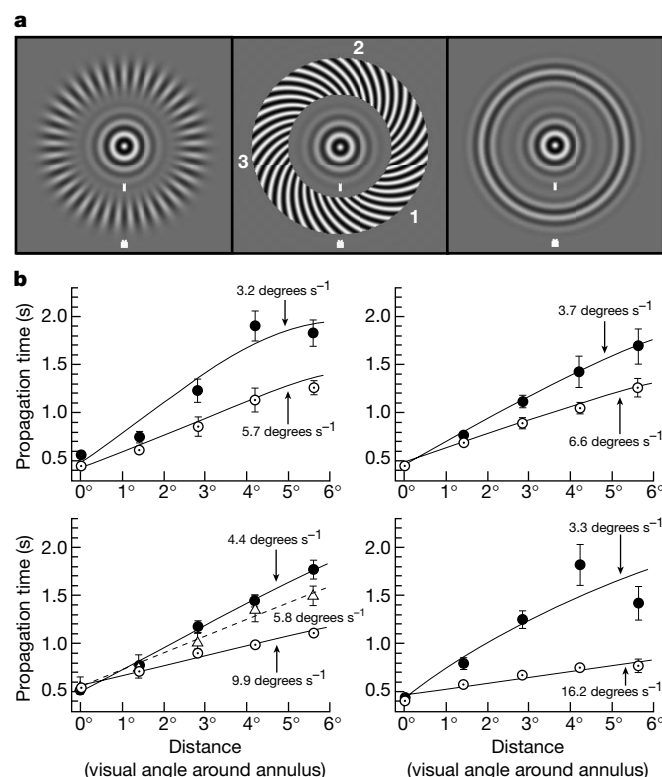


Figure 1 Rivalry stimuli and data of propagation times for dominance waves. **a**, In all experiments one eye viewed the high-contrast spiral grating (middle), while the other eye viewed either the lower-contrast radial (left) or concentric grating (right). Viewers can experience dominance wave propagation by free-fusing the bull's-eyes. (Anaglyphic versions of these stimuli and demonstrations of triggering are available at <http://www.psy.vanderbilt.edu/faculty/blake/rivalry/waves.html>). Typically, when the radial grating is pitted in rivalry against the spiral, one small portion of radial grating achieves local dominance, and this propagates around the annulus. **b**, Propagation times for four observers (H.R.W., top left; R.B., top right; S.L., bottom left; K.S., bottom right) as a function of distance in degrees of visual angle around the annulus. Propagation times were significantly longer for the radial grating (filled circles) than for the concentric grating (open circles), and times for the spiral grating (open triangles) were intermediate. Lines are the best fits of equation (1), and standard errors are indicated.

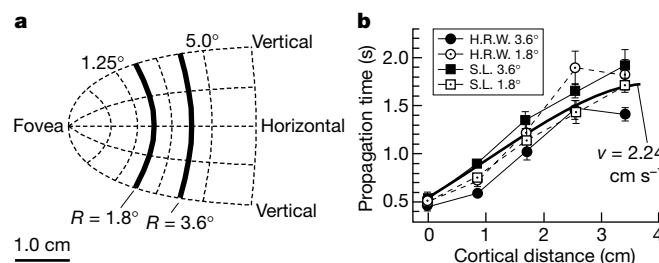


Figure 2 Dependence of propagation times on cortical distance. **a**, Best-fitting complex, logarithmic approximation (dashed lines) to a flattened retinotopic map of human V1 reported previously¹⁷. Thick lines plot the mapping of half annuli with radii of 1.8 and 3.6° . Distance around the annulus was converted into centimetres across cortex using the formulae: $1.0^\circ = 0.6$ cm (1.8° radius); $1.0^\circ = 0.3$ cm (3.6° radius). **b**, Radial pattern data for two subjects and two eccentricities indicate that propagation times are roughly constant in cortical coordinates. The best fit of equation (1) (thick line) produced an estimate of cortical speed of 2.24 cm s^{-1} .

pattern were gathered on two observers under these conditions. For both of the observers, speed derived from equation (1) increased substantially: to $8.29 \text{ degrees s}^{-1}$ for H.R.W.; $6.36 \text{ degrees s}^{-1}$ for S.L. Using the hypothesis that dominance waves might be propagating across primary visual cortex (V1), we converted all data into a graph of propagation time versus cortical distance (in centimetres) (Fig. 2) on the basis of a detailed surface map of human V1 (ref. 17) (Fig. 2a). Data plotted in cortical coordinates collapse to approximate a single function corresponding to a mean propagation speed of 2.24 cm s^{-1} across V1.

As these experiments used an arrival point on the vertical meridian, measured times would not have involved propagation across the corpus callosum. To determine whether inter-hemispheric transfer might slow wave propagation, we shifted the arrival point to one of the lower diagonals (Fig. 1a, point 1), and triggered waves at one of two equidistant points 125° around the annulus (Fig. 1a, points 2 and 3). (The mirror reflection of this configuration produced identical results.) Thus, the shortest path of waves from point 2 to point 1 remained within a hemisphere, whereas waves initiated at point 3 had to traverse the corpus callosum before arriving at point 1. Repeated, randomly interleaved measurements of the propagation times along these two paths revealed that propagation along the intra-hemispheric path was indeed faster ($1.66 \pm 0.64 \text{ s}$) than propagation along the path requiring callosal transfer ($1.84 \pm 0.58 \text{ s}$), and this difference was statistically significant ($t_{254} = 2.24$; $P < 0.02$). Propagation of dominance waves across the corpus callosum thus entails a small time penalty averaging 173 ms, presumably reflecting both callosal transit and time to rekindle a wave in the opposite hemisphere.

The simple cortical network model shown in Fig. 3a can explain dominance wave propagation. Two layers of cortical neurons, one sensitive to the spiral pattern (S) and the other to the target pattern (T), comprise the model. Through interneurons (I_S and I_T) each neuron in each layer inhibits a range of neurons in the opposing layer. The spatial range of inhibition extended to adjacent ocular dominance columns on the basis of human cortical data¹⁸. To simulate collinear facilitation, weak reciprocal excitatory connections were included among neighbouring neurons. To permit dominance alternations, the excitatory neurons in both layers

(but not the inhibitory neurons) incorporated slow, hyperpolarizing currents to produce spike-rate adaptation¹⁹. Neural responses were described by 'spike-rate' equations developed elsewhere^{20,21} (see Methods). This model is similar to a recent model for V1 functional connectivity²².

The model was simulated as a ring of T, S, I_T and I_S neurons ($n = 136$ each) spaced 0.5-mm apart. Time constants were chosen so that the model would produce pattern superposition for the first 150 ms of stimulation, thus replicating results on human rivalry²³. Once the S neurons were dominant, a local pulse to the suppressed T neurons triggered a wave that spread at constant speed around the ring (Fig. 3b). Simulations of the radial pattern without collinear facilitation produced a dominance wave speed of 2.24 cm s^{-1} . Collinear facilitation was introduced for concentric patterns, and this increased wave speed to 4.40 cm s^{-1} . Spiral targets with significantly reduced collinear facilitation produced a speed of 2.68 cm s^{-1} . Furthermore, simulation of radial target patterns with gaps showed that model dominance waves were blocked by a three-cycle-wide gap but could jump a one-cycle gap. Thus, all simulation results agree with our data.

This model goes beyond previous rivalry models^{24,25} by providing a neural mechanism for dominance wave propagation and in incorporating collinear facilitation, which has been shown to enhance dominance¹⁶. For radial gratings, dominance wave propagation is generated by recurrent disinhibition with a spatial spread approximating the distance between human ocular-dominance columns¹⁸. It is known that single-layer recurrent inhibitory networks can support wave propagation⁵, so our network provides a further mechanism for inhibitory wave propagation. To obtain the twofold speed increase for collinear patterns in the model, it was necessary to extend the spatial range of recurrent excitation to twice that of the reciprocal inhibition; more localized recurrent excitation produced much smaller speed increases as exemplified by the spiral simulations. Context-dependent effects in V1 have been explained by similar long-range collinear excitation²².

Our data clearly implicate a retinotopically organized visual area as the site of dominance wave propagation. Our data, however, do not rule out the involvement in rivalry of higher levels of the ventral visual stream as suggested by several studies^{26–28}. Human functional

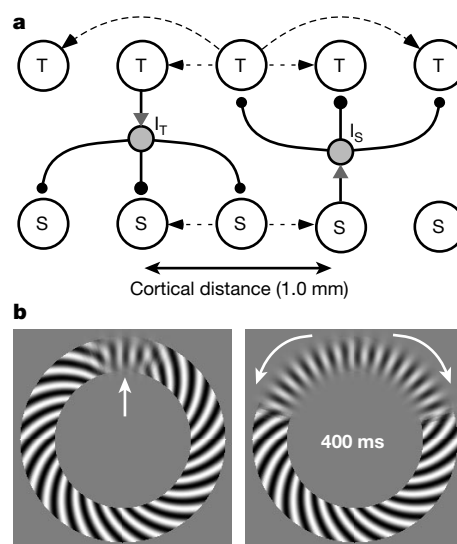


Figure 3 Neural model for dominance wave propagation. **a**, Units driven by the spiral (S) or target (T) monocular pattern (radial, concentric, or spiral) generate mutual inhibition driven by interneurons (grey circles I_S and I_T). The spatial spread of this inhibition produces dominance wave propagation through recurrent disinhibition. Collinear facilitation was simulated through recurrent excitation (dashed arrows). **b**, Dominance wave

propagation generated by equations (2–4) (see Methods). Left, radial pattern dominance has just been triggered locally by a contrast pulse at the arrow. After 400 ms (right) the dominance wave has propagated over 40% of the annulus (arrows indicate propagation direction).

magnetic resonance imaging has revealed signatures of rivalry both in V1 (ref. 29) and in the extrastriate fusiform face area³⁰. One possible resolution is that rivalry may be mediated by a complex interplay involving several cortical areas. In this scheme rivalry waves reflect activity propagation across V1, while more complex stimulus^{28,30} or spatio-temporal conditions²⁷ accentuate aspects of rivalry dependent on higher cortical levels. In any case, our data demonstrate that the site of dominance wave propagation is retinotopically organized, has cortical magnification similar to V1, and exhibits collinear facilitation.

Our technique for measuring the speed of dominance wave propagation should prove useful in future studies of rivalry and other dynamic cortical phenomena. For example, the technique might be applied to behaving primates using spatial electrode arrays either in V1 or higher centres where rivalry has been reported. In addition, the enhanced speed of propagation along collinear contours provides a window into cortical excitability, as demonstrated by our model. It is tempting to suppose that individual and age-related differences in rivalry alternation rates may reflect differences in such collinear facilitation. These and other possibilities can now be tested using dominance wave speed as a sensitive probe. □

Methods

Experiments

Four observers (one naive) participated in one or more of these experiments. Annular rival patterns were generated on a 21-inch NEC monitor (1,024 × 768 resolution; 100 Hz frame rate) controlled by a Power Macintosh computer. The two annular patterns, each 1.8° in mean radius (except for the eccentricity experiments), were viewed through a mirror stereoscope with the head stabilized by a chin and head rest. Width of the spiral annulus was 0.93°, while the gaussian half-width of the radial or concentric pattern annulus was 0.46°. The spiral cosine pattern had a pitch angle of 45° for the high-contrast mask so as to be at the same relative orientation with respect to both radial and concentric annular patterns (Fig. 1a). (Spiral target patterns had a -45° pitch so as to be locally orthogonal to the spiral mask.) Radial and concentric cosine grating contrast was adjusted for each observer to a value at which the spiral (100% contrast) was completely dominant for most of the viewing period—grating contrast varied among observers from 15–25%.

During the experiment, observers maintained strict fixation on the identical bull's-eye fusion targets in the centre of the pattern. Once the low-contrast grating was completely suppressed, the spiral alone being dominant, the observer depressed a switch. This produced a 100-ms contrast increment in the suppressed grating at one of eight equally spaced cardinal locations. The increment comprised a gaussian spatial envelope with half-width along the annulus of 18 arcmin; the magnitude of the contrast increment was 70%, a value sufficient to penetrate suppression locally on every trial. With fixation always maintained on the bull's-eye, the observer monitored the phenomenal status of the grating region, demarcated by two nonius lines. Once that portion of the grating became dominant, the observer released the switch, thereby recording transit time to the arrival point. Trials were run in blocks of 32 with rest periods as needed.

Model

Model simulations were conducted in Matlab software on a Macintosh G4 computer using a Runge–Kutta routine with constant step size (0.25 ms). The equations are^{20,21}:

$$\tau \frac{dT_n}{dt} = -T_n + \frac{100P_+^2}{(10 + H_{T_n})^2 + P_+^2} \quad \text{where } \tau = 20 \text{ ms}$$

and

$$P = E_T - 0.27 \sum_k I_{sk} \exp(-x_{nk}^2/\sigma^2) + g \underbrace{\sum_{k \neq n} T_k \exp(-x_{nk}^2/(2\sigma)^2)}_{\text{collinear facilitation term}} \quad (2)$$

$$\tau_I \frac{dI_{Tn}}{dt} = -I_{Tn} + T_n \quad \text{where } \tau_I = 11 \text{ ms} \quad (3)$$

$$\tau_H \frac{dH_{Tn}}{dt} = -H_{Tn} + 2T_n \quad \text{where } \tau_H = 900 \text{ ms} \quad (4)$$

where T_n is the firing rate of an excitatory neuron driven by the target stimulus, I_{sk} and I_{Tn} are firing rates of inhibitory neurons driven by the respective excitatory neurons, and H_{Tn} is the spike frequency adaptation variable for T_n . The input P to each T neuron includes a constant E_T representing the stimulus strength, subtractive inhibition from a spatially weighted sum of I_{sk} cells ($\sigma = 1.0$ mm), and a term embodying collinear facilitation. The 900-ms time constant for the dH/dt equation is based on slow after-hyperpolarizing

(AHP) potentials in human excitatory neurons¹⁹.

A comparable set of equations describes activity of S neurons driven by the spiral. For concentric patterns, the collinear facilitation parameter $g = 0.04$, but $g = 0$ for radial patterns. The intermediate case of spirals was approximated by $g = 0.02$ and reducing the spatial spread of excitation by half. Parameters were chosen to reflect available anatomical and physiological data. Thus, the time constant for inhibitory neurons (described by a linear equation for simplicity) was faster than that for excitatory cells, reflecting properties of cortical fast-spiking neurons. In equation (2) $P_+ = \max(P, 0)$ so that negative inputs drive the firing rate to zero. The maximum firing rate of the S and T neurons was chosen to be 100 by convention, and the Naka–Rushton nonlinearity has been related to cortical physiology elsewhere^{20,21}. Effective stimulus strengths of the rivaling monocular patterns were chosen to be $E_S = 30$ for the high contrast spiral grating and $E_T = 24$ for the lower-contrast target stimulus. Variations of these numbers changed wave propagation speed by no more than 10%. Maximum strength of collinear excitation was constrained, using standard analytical techniques, so there could be no uniformly excited equilibrium state in the network²⁰. Although our model is deterministic, addition of noise can easily generate a gamma function distribution of intervals²⁴.

Received 6 April; accepted 1 July 2001.

1. Winfree, A. T. *When Time Breaks Down: The Three-Dimensional Dynamics of Electrochemical Waves and Cardiac Arrhythmias*. (Princeton Univ. Press, New Jersey, 1987).
2. Nagai, Y., González, H., Shrier, A. & Glass, L. Paroxysmal starting and stopping of circulating waves in excitable media. *Phys. Rev. Lett.* **84**, 4248–4251 (2000).
3. Basarsky, T. A., Duffy, S. N., Andrew, R. D. & MacVicar, B. A. Imaging spreading depression and associated intracellular calcium waves in brain slices. *J. Neurosci.* **18**, 7189–7199 (1998).
4. Kim, U., Bal, T. & McCormick, D. A. Spindle waves are propagating synchronized oscillations in the ferret LGNd in vitro. *J. Neurophysiol.* **74**, 1301–1323 (1995).
5. Rinzel, J., Terman, D., Wang, X. J. & Ermentrout, B. Propagating activity patterns in large scale inhibitory neuronal networks. *Science* **279**, 1351–1355 (1998).
6. Ermentrout, G. B. & Kleinfeld, D. Traveling electrical waves in cortex: insights from phase dynamics and speculation on a computational role. *Neuron* **29**, 33–44 (2001).
7. Crick, F. & Koch, C. Consciousness and neuroscience. *Cerebral Cortex* **8**, 97–107 (1998).
8. Logothetis, N. K. Single units and conscious vision. *Phil. Trans. R. Soc. Lond. B* **353**, 1801–1818 (1998).
9. Blake, R., O'Shea, R. P. & Mueller, T. J. Spatial zones of binocular rivalry in central and peripheral vision. *Vis. Neurosci.* **8**, 469–478 (1992).
10. Blake, R., Westendorff, D. & Fox, R. Temporal perturbations of binocular rivalry. *Percept. Psychophys.* **48**, 593–602 (1990).
11. Walker, P. & Powell, D. J. The sensitivity of binocular rivalry to changes in the nondominant stimulus. *Vision Res.* **19**, 247–249 (1979).
12. Malach, R., Amir, Y., Harel, M. & Grinvald, A. Relationship between intrinsic connections and functional architecture revealed by optical imaging and *in vivo* targeted biocytin injections in primate striate cortex. *Proc. Natl Acad. Sci. USA* **90**, 10469–10473 (1993).
13. Das, A. & Gilbert, C. D. Long-range horizontal connections and their role in cortical reorganization revealed by optical recording of cat primary visual cortex. *Nature* **375**, 780–784 (1995).
14. Field, D. J., Hayes, A. & Hess, R. F. Contour integrations by the human visual system: evidence for a local 'association' field. *Vision Res.* **33**, 173–193 (1993).
15. Kamitani, Y. & Shimojo, S. Manifestation of scotomas created by transcranial magnetic stimulation of human visual cortex. *Nature Neurosci.* **2**, 767–771 (1999).
16. Alais, D. & Blake, R. Grouping visual features during binocular rivalry. *Vision Res.* **39**, 4341–4353 (1999).
17. Horton, J. C. & Hoyt, W. F. The representation of the visual field in human striate cortex: a revision of the classic Holmes map. *Arch. Ophthalmol.* **109**, 816–824 (1991).
18. Hitchcock, P. F. & Hickey, T. L. Ocular dominance columns: evidence for their presence in humans. *Brain Res.* **182**, 176–179 (1980).
19. McCormick, D. A. & Williamson, A. Convergence and divergence of neurotransmitter action in human cerebral cortex. *Proc. Natl Acad. Sci. USA* **86**, 8098–8102 (1989).
20. Wilson, H. R. *Spikes, Decisions, and Actions: Dynamical Foundations of Neuroscience*. (Oxford Univ. Press, Oxford, 1999).
21. Wilson, H. R., Krupa, B. & Wilkinson, F. Dynamics of perceptual oscillations in form vision. *Nature Neurosci.* **3**, 170–176 (2000).
22. Somers, D. C. et al. A local circuit approach to understanding integration of long range inputs in primary visual cortex. *Cereb. Cort.* **8**, 204–217 (1998).
23. Wolfe, J. M. Influence of spatial frequency, luminance, and duration on binocular rivalry and abnormal fusion of briefly presented dichoptic stimuli. *Perception* **12**, 447–456 (1983).
24. Lehky, S. An stable multivibrator model of binocular rivalry. *Perception* **17**, 215–228 (1988).
25. Blake, R. A neural theory of binocular rivalry. *Psych. Rev.* **96**, 145–167 (1989).
26. Leopold, D. A. & Logothetis, N. K. Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature* **379**, 549–553 (1996).
27. Logothetis, N. K., Leopold, D. A. & Sheinberg, D. L. What is rivaling during binocular rivalry? *Nature* **380**, 621–624 (1996).
28. Kovács, I., Papathomas, T. V., Yang, M. & Fehér, A. When the brain changes its mind: interocular grouping during binocular rivalry. *Proc. Natl Acad. Sci. USA* **93**, 15508–15511 (1996).
29. Polonsky, A., Blake, R., Braun, J. & Heeger, D. Neuronal activity in human primary visual cortex correlates with perception during binocular rivalry. *Nature Neurosci.* **3**, 1153–1159 (2000).
30. Tong, F., Nakayama, K., Vaughan, J. T. & Kanwisher, N. Binocular rivalry and visual awareness in human extrastriate cortex. *Neuron* **21**, 753–759 (1998).

Acknowledgements

This research was supported by grants from NSERC (H.R.W.) and NIH (R.B.).

Correspondence and requests for materials should be addressed to H.R.W. (email: hrwilson@yorku.ca).

Two-step process for photoreceptor formation in *Drosophila*

Bertrand Mollereau^{*,†}, Maria Dominguez^{†,‡}, Rebecca Webel[§], Nansi Jo Colley[§], Benison Keung^{*}, Jose F. de Celis^{||} & Claude Desplan^{*}

^{*} Department of Biology, New York University, New York, USA

[‡] Instituto de Neurociencias, UMH-CSIC, Campus de San Juan Alicante 03550, Spain

[§] Departments of Ophthalmology and Visual Science and Genetics, University of Wisconsin, Madison, Wisconsin 53792, USA

^{||} Centro de Biología Molecular 'Severo Ochoa', Universidad Autónoma de Madrid, Madrid 28049, Spain

[†] These authors contributed equally to this work

The formation of photoreceptor cells (PRCs) in *Drosophila* serves as a paradigm for understanding neuronal determination and differentiation. During larval stages, a precise series of sequential inductive processes leads to the recruitment of eight distinct PRCs (R1–R8)¹. But, final photoreceptor differentiation, including rhabdomere morphogenesis and *opsin* expression, is completed four days later, during pupal development^{2,3}. It is thought that photoreceptor cell fate is irreversibly established during larval development, when each photoreceptor expresses a particular set of transcriptional regulators and sends its projection to different layers of the optic lobes. Here, we show that the *spalt* (*sal*) gene complex^{4–7} encodes two transcription factors that are required late in pupation for photoreceptor differentiation. In the absence of the *sal* complex, rhabdomere morphology and expression of *opsin* genes in the inner PRCs R7 and R8 are changed to become identical to those of outer R1–R6 PRCs. However, these cells maintain their normal projections to the medulla part of the optic lobe, and not to the lamina where outer PRCs project. These data indicate that photoreceptor differentiation occurs as a two-step process. First, during larval development, the photoreceptor neurons become committed and send their axonal projections

to their targets in the brain. Second, terminal differentiation is executed during pupal development and the photoreceptors adopt their final cellular properties.

The two zinc finger proteins of the *sal* gene complex are expressed in distinct subsets of PRCs throughout eye development^{8,9}. *sal major* (*salm*) and *sal related* (*salr*) have almost identical expression patterns in most tissues, including the imaginal disc, and thus are likely to have similar or overlapping roles. In eye imaginal discs, the *sal* genes are expressed in a very dynamic pattern including R3 and R4 PRCs and cone cells. However, no obvious defects have been reported in *salm* mutant eye discs (U. Gaul and M. Mlodzik, personal communications). In the adult, *salm* is no longer expressed in outer PRCs but is restricted to the inner PRCs, R7 and R8 (ref. 9). To determine when this transition happens, we analysed *salm* expression during pupal life. After 24 h of pupation, *salm* was expressed in R3, R4 and cone cells (Fig. 1a and b). After 48 h, *salm* expression was strongly diminished in the cone cells with only weak labelling detected at 72 h in these cells (Fig. 1c and d). Between 48 h and 60 h of pupation, *salm* expression was turned off in R3 and

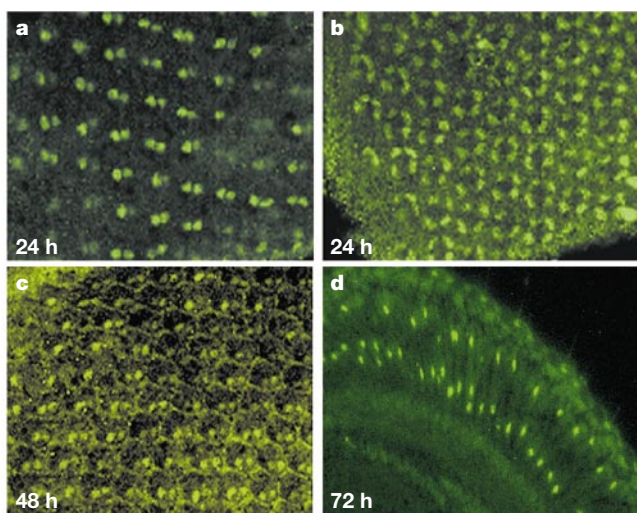


Figure 1 Expression of Salm during eye pupal maturation. Twenty-four hour pupal retina showing Salm expression in R3 and R4 (a) and in cone cells (b). Cone cells are located in a focal plane above R3 and R4. After 48 h, Salm expression is lost in cone cells but remains in R3 and R4 (c). After 72 hours, Salm expression is limited to R7 and R8 where it remains to adulthood. Weak expression of Salm is also detected in cone cells at 72 h (d). R7 cells were identified with anti-Prospero antibodies (shown in Fig. 4).

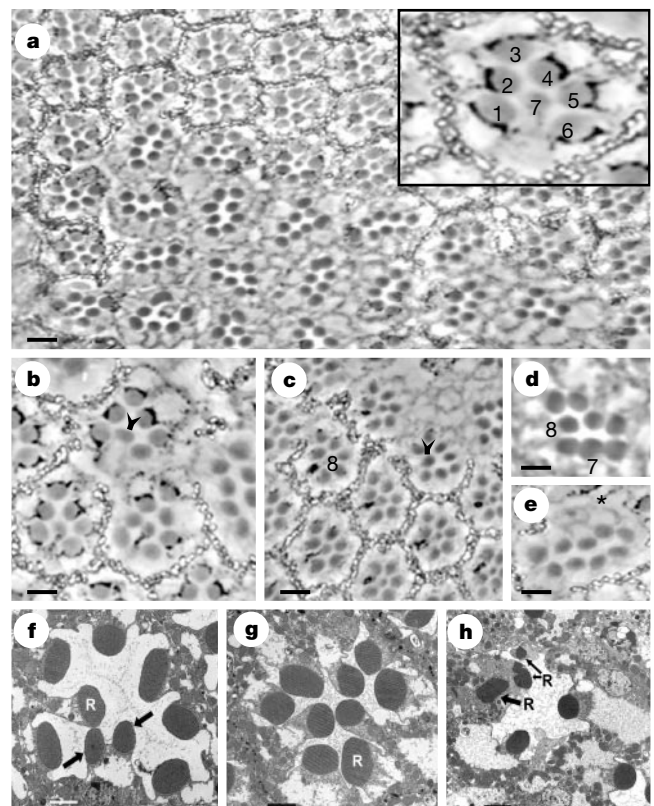


Figure 2 *salm/salr* mutant clones reveal transformation of inner photoreceptor cells (PRCs) into outer PRCs. a–e, Phase-contrast images of tangential sections through mosaic *salm/salr* mutant eye tissue. Wild-type cells are pigmented (close-up in a) and *salm/salr* cells are unpigmented. Whenever an R7 (arrowhead in b) or R8 (arrowhead in c) cells is mutant, it forms a large rhabdomere typical of outer PRCs. Typically, mutant ommatidia lack both R7 and R8 and have a concomitant gain of outer PRCs (d). In addition, a small proportion of mutant ommatidia have additional outer PRCs (asterisk in e). Scale bars, 5 µm (a, c); 4 µm (b); 3.3 µm (d, e). f–h, Electron microscopy analysis of *salm/salr* mutant ommatidia: *ey-FLP GMR hid CL/F40Df(2L)5*. One-week-old *salm/salr* mutant ommatidia contains a photoreceptor with duplicated rhabdomeres (arrows in f) (scale bar, 1.5 µm), or display nine individual rhabdomeres connected to separate cell bodies (scale bar, 2.0 µm) (g). Eight week-old *salm/salr* mutant ommatidium containing PRCs with reduced rhabdomeres or internalized rhabdomeres (arrows in h) (scale bar, 2.2 µm), as well as cells lacking rhabdomeres. R, rhabdomere. Numbers 1–8 stand for photoreceptor cell numbers, defined by their position and size.

R4 but was activated in R7 and R8 where it was maintained throughout adult life (Fig. 1d and ref. 9). This activation of *salm* expression in inner PRCs occurred at about the same time as the onset of rhabdomere morphogenesis and *rhodopsin* expression^{10,11}, suggesting a role for these genes in the differentiation of R7 and R8. To examine the role of the *sal* complex in eye morphogenesis, we generated tissue that was mutant for both *salm* and *salr* using a small chromosomal deficiency (Df(2L)32EP5) uncovering only these genes⁸ (Fig. 2). No large phenotypic changes were detected in the PRCs in imaginal discs (not shown) and the projections to the optic lobes appeared to be normal. However, the mutant ommatidia in the adult eye were greatly altered: the small central rhabdomeres of inner R7 and R8 PRCs were absent, and extra PRCs with large rhabdomeres were observed (Fig. 2a). Mosaic analysis indicated that the individual mutant R7 and R8 cells exhibited features of outer R1–R6 PRCs in a cell-autonomous manner (Fig. 2b and c). Sections through mutant ommatidia showed that most of the transformed PRCs had rhabdomeres that extended throughout the thickness of the retina, and hence ommatidia with eight outer PRCs were found in apical sections (Figs 2d and 3b). Together these data suggest that both R7 and R8 are transformed into outer PRCs. In addition, we also observed a small proportion of ommatidia with as many as nine or ten outer photoreceptor rhabdomeres (Figs 2e and 3b), which appear to arise both from rhabdomere duplication (Fig. 2f) and from recruitment of additional PRCs, on the basis of the presence of extra photoreceptor cell bodies in some ommatidia (Fig. 2g). Because of the position of this extra rhabdomere between R3 and

R4 cells in most cases (not shown), we favour the transformation of ‘mystery cells’ toward outer photoreceptor cell fate¹². Finally, in a few cases, six or less PRCs were present. This could be accounted for by photoreceptor loss: older flies exhibited dramatic pathology of some PRCs reminiscent of degeneration (Fig. 2h).

In wild-type flies, the six outer PRCs, R1–R6, express *rhodopsin1* (*rh1*) and mediate image formation and dim light vision. The inner PRCs R7 and R8 express distinct *rhodopsins* (*rh3* or *rh4* in R7; *rh5* or *rh6* in R8) that mediate perception^{13–15}. To define more precisely the cell fates adopted by *sal* mutant ommatidia, we examined *rhodopsin* expression in eyes that were completely mutant for *salm/salr*¹⁶ (Fig. 3). All rhabdomeres contained Rh1 (Fig. 3b) but none of the inner rhodopsins (Rh3, Rh4, Rh5 and Rh6) were detected at significant levels (Fig. 3d, f and h), strongly supporting the model proposed above that both R7 and R8 were transformed into outer R1–R6 PRCs. A similar, but weaker and less penetrant phenotype was observed with a single mutant for *salm* (*salm*⁶⁵ allele) (data not shown), probably owing to a partially redundant function for this family of related transcription factors. Because *salm* is normally only expressed in non-Rh1-positive PRCs, and because *rh1* expression is expanded in *salm/salr* mutants, it is possible that *rh1* is repressed by *salm/salr*. However, the extent to which *Sal* proteins regulate *rhodopsin* promoters is as yet unknown.

The fairly late timing of *salm* expression in inner PRCs during pupal life (that is, much later than the time PRCs send out their axons in third instar larvae) suggested that its role in photoreceptor differentiation was not related to early photoreceptor specification or axon pathfinding. Consistent with this hypothesis, we found that, in the *salm/salr* mutants, the transformed R7 and R8 cells projected their axons to the medulla, which is their normal site of projection (Fig. 4b and d). Therefore, it appears that the determination of R7 and R8 is correctly initiated, but that these cells later adopt features typical of outer R1–R6 PRCs. Furthermore, the expression of *prospero* (*pros*), an early cell-marker for R7 neurons¹⁷ whose expression is maintained in the adult and controls aspects of R7 differentiation (T. Cook and C.D., manuscript in preparation; and

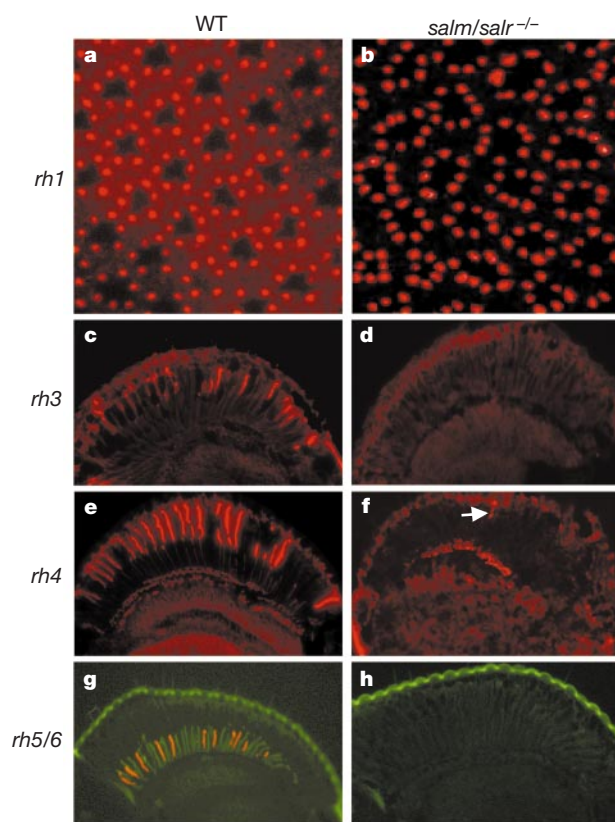


Figure 3 Expression of rhodopsins is altered in *salm/salr* (Df(2L)5) mutant eyes. The use of an adaptation of *EGUF/hid* technique allows the generation of eyes completely mutant for *salm/salr* (b, d, f, h). Fluorescent anti-Rh1 staining in tangential eye sections in wild-type (WT) (a) and *salm/salr* mutants (b). Anti-Rh3, anti-Rh4 or anti-Rh5/6 stainings in WT (c, e, g), and *salm/salr* mutant (d, f, h). Rh5 is stained in red and Rh6 in green. Rh1 is expanded to inner PRCs (b), whereas Rh3, Rh5 and Rh6 are completely lost (d, h) and very little Rh4 remains (arrowhead in f).

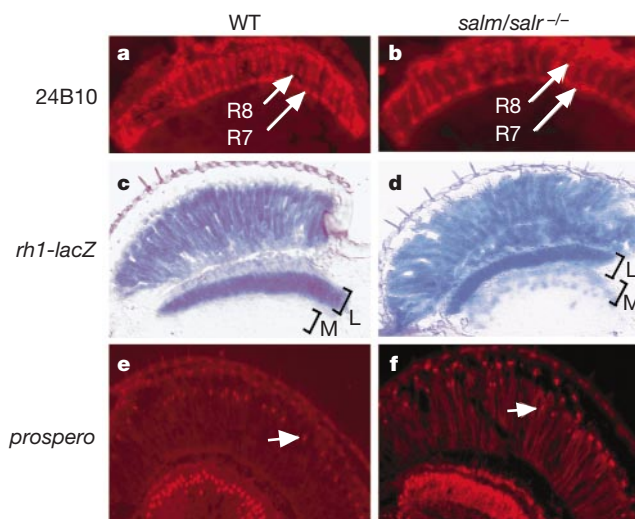


Figure 4 R7 and R8 projections and *pros* expression remain unaltered in *salm/salr* mutant eyes. R7 and R8 projections to the medulla are revealed with 24B10 antibodies in WT (a) and *salm/salr* mutants (b). The two levels of projections of R7 and R8 are distinguishable (arrows). (c), In WT, *rh1-lacZ* staining spans the whole retina and outer PRCs project only to the lamina. (d), In *salm/salr* mutants, *rh1-lacZ* reveals additional projections to the medulla, demonstrating that R7 and R8 express *rh1* but still project to the medulla. L, lamina; M, Medulla. *Prospero*, an early R7 marker, is present in R7 nuclei in WT (e) and remains in transformed R7 cells in *salm/salr* mutants (f). Staining in the upper part of the retina reveals the cone cells.

Fig. 4e), is normal in adult clones of *salm/salr* (Fig. 4f). These results demonstrate that although the mutant R7 and R8 cells have the morphology of outer PRCs and express *rh1*, they express early R7-specific markers. There are examples of transformation of inner into outer PRCs, for instance in mis-expression experiments with *rough* or *seven-up*^{18,19}. However, in these cases, the transformation occurs much earlier in the disc and concerns only R7 cells. Although this has not been addressed, we predict that, in this case, the projections of transformed R7 cells are to the lamina and no longer to the medulla.

Together, our data show that the *sal* complex is essential for the terminal differentiation of inner R7 and R8 PRCs. In its absence, these PRCs exhibit characteristics of both inner and outer PRCs. Although initial work had predicted that eye development occurs in different stages marked by the successive expression of various molecules in PRCs²⁰, most recent studies largely assumed that PRCs are fully determined in the third instar imaginal disc²¹. Here, we demonstrate that photoreceptor development is a two-step process and that each step is under different genetic regulation. In the first step, the cells adopt their fate as neurons, become committed, and send specific axonal projections. During this recruitment stage, the PRCs are predetermined but their fate is not fully and irreversibly established. In a second step these neurons become mature photoreceptors. They execute their differentiation program and acquire their final properties with *rhodopsin* gene expression and rhabdomere morphogenesis^{10,22–24}. Atypical terminal differentiation of inner PRCs occurs naturally in specific parts of the retina. For instance, two rows of ommatidia at the dorsal margin of the eye display normal R7 and R8 specification but later acquire different terminal fates, with much larger rhabdomeres and cells that express only *rh3* in both R7 and R8 (ref. 25). These cells still project normally to the medulla. We note that in *Crx*-deficient mice, a model for human cone-rod dystrophy, the photoreceptors are specified but fail to undergo terminal differentiation, with no outer segment morphogenesis and loss of cone and rod opsins²⁶. □

Methods

Mosaic analysis

To generate marked *salm/salr* homozygous mutant clones in the eye, *w; FRT40 Df(2L)32FP-5/Cyo* males⁸ were crossed with *w hs-FLP; FRT40 P[w+30C/Cyo* females. To generate *salm/salr* or *salm* (*salm*⁶⁵) clones in the *GMR-hid* background¹⁶ *FRT40 Df(2L)32FP-5/Cyo* or *FRT40 salm*⁶⁵ males (a gift from U. Gaul) were crossed with *eyFLP; FRT40GMRhid CL* females. The wild-type tissue was eliminated by the combination of the *GMR-hid* transgene and a cell lethal gene, which together induce apoptosis in tissues that are non-homozygous for *salm/salr* or *salm*.

Immunolabelling

Tangential sections of adult eyes²⁷ and fluorescent immunolabelling²⁸, were performed as described in refs 27 and 28. Early pupal eye discs after 24 and 48 h of pupation (about 25% and 50% pupation) were isolated and stained directly, while late-pupal eye discs (60 and 72 h) were frozen, sectioned horizontally and stained as for adult eye sections. Rabbit anti-Rh3 (diluted 1/100) and rabbit anti-Rh4 (1/200) antibodies were provided by C. Zuker. Mouse anti-Rh5 (1/20) antibodies were a gift from S. Britt²⁹. Rabbit anti-Rh6 antibodies (1/5,000) were generated in our laboratory (D. Killian and P. Beaufils). Rat anti-Salm (1/100) was a gift from R. Barrio⁴, and mouse anti-Rh1 antibodies (1/50) and mouse 24B10 antibody (1/20) were from the Developmental Studies Hybridoma Bank. Mouse anti-Pros antibodies (1/4) were provided by C. Doe¹⁷. Antibodies (dilutions are given in parentheses above) were incubated overnight at 4 °C on horizontal or tangential eye sections. After several washes, a secondary FITC goat anti-rabbit antibody (diluted 1/200), or a secondary Cy3 goat anti-mouse or anti-rabbit (diluted 1/400) (Jackson) were incubated for 2 h at room temperature, washed and mounted.

Electron microscopy and ultrastructural analysis

Fixation and processing of adult heads were carried out by a modification of the methods of Baumann and Walz as described previously³⁰. The fixed tissue was dehydrated followed by propylene oxide treatment and embedded in Spurr's medium (Polysciences). Ultrathin sections were produced using a Reichert Ultracut E ultramicrotome. Sections were stained with 2% uranyl acetate and lead citrate and viewed at 80 kV on a Phillips 410 electron microscope. Flies were collected from the beginning of eclosion to seven weeks old. For each time point at least three individual heads were sectioned and 100 ommatidia were observed from each eye. Sections were obtained from apical, mid and basal regions of the eye to assess the phenotype and number of photoreceptor cells throughout the retina.

Received 29 March; accepted 29 June 2001.

- Dominguez, M., Wasserman, J. D. & Freeman, M. Multiple functions of the EGF receptor in *Drosophila* eye development. *Curr. Biol.* **8**, 1039–1048 (1998).
- Perry, M. M. Further studies on the development of the eye of *Drosophila melanogaster*. I. The ommatidia. *J. Morphol.* **124**, 227–248 (1968).
- Cagan, R. L. & Ready, D. F. The emergence of order in the *Drosophila* pupal retina. *Dev. Biol.* **136**, 346–362 (1989).
- Kuhnlein, R. P. *et al.* *Spalt* encodes an evolutionarily conserved zinc finger protein of novel structure which provides homeotic gene function in the head and tail region of the *Drosophila* embryo. *EMBO J.* **13**, 168–179 (1994).
- de Celis, J. F., Barrio, R. & Kafatos, F. C. A gene complex acting downstream of *dpp* in *Drosophila* wing morphogenesis. *Nature* **381**, 421–442 (1996).
- Rusten, T. E. *et al.* *Spalt* modifies EGFR-mediated induction of chordotonal precursors in the embryonic PNS or *Drosophila* promoting the development of oenocytes. *Development* **128**, 711–722 (2001).
- Elstob, P. R., Brodus, V. & Gould, A. P. *Spalt*-dependent switching between two cell fates that are induced by the *Drosophila* EGF receptor. *Development* **128**, 723–732 (2001).
- Barrio, R., de Celis, J. F., Bolshakov, S. & Kafatos, F. C. Identification of regulatory regions driving the expression of the *Drosophila* *spalt* complex at different developmental stages. *Dev. Biol.* **215**, 33–47 (1999).
- Mollereau, B. *et al.* A green fluorescent protein enhancer trap screen in *Drosophila* photoreceptor cells. *Mech. Dev.* **93**, 151–160 (2000).
- Kumar, J. P. & Ready, D. F. Rhodopsin plays an essential structural role in *Drosophila* photoreceptor development. *Development* **121**, 4359–4370 (1995).
- Sheng, G., Thouvenot, E., Schmucker, D., Wilson, D. S. & Desplan, C. Direct regulation of rhodopsin 1 by Pax-6/eyeless in *Drosophila*: evidence for a conserved function in photoreceptors. *Genes Dev.* **11**, 1122–1131 (1997).
- Cadavid, A. L., Ginzel, A. & Fischer, J. A. The function of the *Drosophila* fat facets deubiquitinating enzyme in limiting photoreceptor cell number is intimately associated with endocytosis. *Development* **127**, 1727–1736 (2000).
- Hardie, R. C. in *Sensory Physiology 5* (ed. Ottoson, D.) 1–79 (Springer, Heidelberg, 1985).
- Papatsenko, D., Sheng, G. & Desplan, C. A new rhodopsin in R8 photoreceptors of *Drosophila*; evidence for coordinate expression with Rh3 in R7 cells. *Development* **124**, 1665–1673 (1997).
- Chou, W. H. *et al.* Patterning of the R7 and R8 photoreceptor cells of *Drosophila*: evidence. *Development* **126**, 607–616 (1999).
- Stowers, R. S. & Schwarz, T. L. A genetic method for generating *Drosophila* eyes composed exclusively of mitotic clones of a single genotype. *Genetics* **152**, 1631–1639 (1999).
- Kauffman, R. C., Li, S., Gallagher, P. A., Zhang, J. & Carthew, R. W. Ras1 signaling and transcriptional competence in the R7 cell of *Drosophila*. *Genes Dev.* **10**, 2167–2178 (1996).
- Kimmel, B. E., Heberlein, U. & Rubin, G. M. The homeo domain protein rough is expressed in a subset of cells in the developing *Drosophila* eye where it can specify photoreceptor cell subtype. *Genes Dev.* **4**, 712–727 (1990).
- Begemann, G., Michon, A. M., van der Voorn, L., Wepf, R. & Mlodzik, M. The *Drosophila* orphan nuclear receptor seven-up requires the Ras pathway for its function in photoreceptor determination. *Development* **121**, 225–235 (1995).
- Venkatesh, T. R., Zipursky, S. L. & Benzer, S. Molecular analysis of the development of the compound eye in *Drosophila*. *Trends Neurosci.* **8**, 251–257 (1985).
- Freeman, M. Cell determination strategies in the *Drosophila* eye. *Development* **124**, 261–270 (1997).
- Colley, N. J., Cassill, J. A., Baker, E. K. & Zuker, C. S. Defective intracellular transport is the molecular basis of rhodopsin-dependent dominant retinal degeneration. *Proc. Natl Acad. Sci. USA* **97**, 3070–3074 (1995).
- Kumar, J. P., Bowman, J. O., Touse, J. E. & Ready, D. F. Rhodopsin replacement rescues photoreceptor structure during a critical developmental window. *Dev. Biol.* **188**, 43–47 (1997).
- Chang, H. Y. & Ready, D. F. Rescue of photoreceptor degeneration in rhodopsin-null *Drosophila* mutants by activated Rac1. *Science* **290**, 1978–1980 (2000).
- Fortini, M. E. & Rubin, G. M. The optic lobe projection pattern of polarization-sensitive photoreceptor cells in *Drosophila melanogaster*. *Cell Tissue Res.* **265**, 185–191 (1991).
- Furukawa, T., Morrow, E. M. & Cepko, C. L. *Crx*, a novel *otx*-like homeobox gene, shows photoreceptor-specific expression and regulates photoreceptor differentiation. *Cell* **91**, 531–541 (1997).
- Tomlinson, A. & Ready, D. F. Cell fate in the *Drosophila* ommatidium. *Dev. Biol.* **123**, 264–275 (1987).
- Porter, J. A. & Montell, C. Distinct roles of the *Drosophila* *ninaC* kinase and myosin domains revealed by systematic mutagenesis. *J. Cell Biol.* **122**, 601–612 (1993).
- Chou, W. H. *et al.* Identification of a novel *Drosophila* opsin reveals specific patterning of the R7 and R8 photoreceptor cells. *Neuron* **17**, 1101–1115 (1996).
- Colley, N. J., Baker, E. K., Starnes, M. A. & Zuker, C. S. The cyclophilin homolog *ninaA* is required in the secretory pathway. *Cell* **67**, 255–263 (1991).

Acknowledgements

We are indebted to R. Barrio for her contribution to this work. We thank S. Britt, R. Kuhnlein, C. Doe, C. Zuker, P. Beaufils and K. Basler for fly stocks and antibodies, the Desplan and Treisman laboratories for support and discussion, and I. Tan for help with ultrathin section analysis. We are also grateful to U. Gaul and M. Mlodzik for allowing us to report unpublished results, J. Treisman for support, and T. Cook and F. Pichaud for comments on the manuscript. We would like to thank Developmental Studies Hybridoma Bank for antibodies. B.M. was supported by the Human Frontier Science Program Organization (HFSP). This work was supported by grants from the National Eye Institute (NEI) to C.D., from HHMI, Research to Prevent Blindness (RPB) and the Retina Research Foundation to N.J.C., and the Dirección General de Investigación from Ministerio de Ciencia y Tecnología (MCYT) to M.D.

Correspondence and requests for materials should be addressed to C.D. (e-mail: claudedesplan@nvu.edu) or J.F.D.C. (e-mail: jfcdcelis@trasto.cbm.uam.es).

Virus-mediated killing of cells that lack p53 activity

Kenneth Raj, Phyllis Ogston & Peter Beard

Swiss Institute for Experimental Cancer Research (ISREC),
155, Ch. des Boveresses, 1066 Epalinges, Switzerland

A major goal of molecular oncology is to identify means to kill cells lacking p53 function. Most current cancer therapy is based on damaging cellular DNA by irradiation or chemicals. Recent reports^{1,2} support the notion that, in the event of DNA damage, the p53 tumour-suppressor protein is able to prevent cell death by sustaining an arrest of the cell cycle at the G2 phase. We report here that adeno-associated virus (AAV) selectively induces apoptosis in cells that lack active p53. Cells with intact p53 activity are not killed but undergo arrest in the G2 phase of the cell cycle. This arrest is characterized by an increase in p53 activity and p21 levels and by the targeted destruction of CDC25C. Neither cell killing nor arrest depends upon AAV-encoded proteins. Rather, AAV DNA, which is single-stranded with hairpin structures at both ends^{3,4}, elicits in cells a DNA damage response that, in the absence of active p53, leads to cell death. AAV inhibits tumour growth in mice. Thus viruses can be used to deliver DNA of unusual structure into cells to trigger a DNA damage response without damaging cellular DNA and to selectively eliminate those cells lacking p53 activity.

Saos-2 cells (a p53-null, pRb-null osteosarcoma line) and U2OS cells (which are wild type for p53 and pRb) were infected with AAV-2. Saos-2 cells died, while U2OS cells did not die but enlarged (see Supplementary Information). To exclude replication of the AAV, ultraviolet-inactivated viruses were tested and found to give similar results. Ultraviolet-inactivated AAV was used for all subsequent experiments. Flow cytometry revealed that Saos-2 cells, when infected with AAV, accumulated briefly with 4n DNA content. Cell death occurred soon after (Fig. 1a). Many of these cells contained a quantity of DNA less than 2n (Fig. 1a), which together with Annexin V staining (data not shown) showed that they underwent apoptosis. However, most infected U2OS cells arrested with 4n DNA content and re-entered the cell cycle after several days (Fig. 1b). A minority of cells arrested with 2n DNA content. When higher amounts of AAV were used, most of the U2OS cells arrested

for a prolonged period (Fig. 1c). Whereas lower amounts of virus were sufficient to kill Saos-2 or transiently arrest U2OS cells (Fig. 1a and b), higher amounts were used to prolong the arrest and facilitate analysis of the arrested cells. To investigate the contribution of p53 activity in the response of cells to AAV infection, we inactivated the p53 protein in U2OS by ectopically overexpressing p53DD, a trans-dominant negative p53 mutant⁵. Infection of these cells with AAV resulted in a transient G2 arrest followed by cell death, as seen with Saos-2 cells (Fig. 1d).

When normal human osteoblasts (NHOs) were infected with AAV, they also enlarged and arrested with 4n DNA content (Fig. 1e). When the p53 protein in NHOs was degraded by expression of the human papilloma virus 16 (HPV-16) E6 protein before infection with AAV, the cells arrested at G2 for a short period before dying (Fig. 1f). These observations show that the effect of AAV is not unique to osteosarcomas but is also observable in normal osteoblasts, and that the outcome of this infection is dependent on p53 activity. Consistent with this, the p53 protein in U2OS was stabilized after AAV infection (Fig. 2a). A similar increase was also observed for the p21 protein (Fig. 2a), which is indicative of an increase in p53 activity⁶.

To analyse further the cell cycle block imposed by AAV, we assayed the activity of the cyclin B-cdc2 kinase complex, which triggers the G2/M transition of the cell cycle⁷. Cells blocked in mitosis by nocodazole exhibited high cyclin B-cdc2 kinase activity (Fig. 2b). However, AAV-infected U2OS and Saos-2 cells, despite having a 4n DNA content, possessed cyclin B-cdc2 kinase activity that was even lower than that of the unsynchronized control population, indicating that the AAV-induced block was at the G2 phase. Following infection of U2OS with AAV, a substantial fraction of the cdc2 protein migrated on gel electrophoresis at a slower rate than the control, indicating that it might be phosphorylated (Fig. 2c). Cells treated with nocodazole did not show the reduced migration of cdc2 protein, further suggesting that AAV does not induce an arrest in mitosis. Upon checking the protein level of CDC25C, a phosphatase crucial for activating cdc2 (ref. 8), we found that it decreased dramatically in AAV-infected U2OS (Fig. 2d) but not in AAV-infected U2OS+p53DD cells (Fig. 2e). This disappearance of CDC25C protein was prevented by N-acetyl-leu-leu-norleucinal (NaLLN)⁹, indicating involvement of the proteasome complex (Fig. 2f). This degradation was specific as the level of CDC25B protein (Fig. 2d) and that of many other proteins tested were unchanged.

To determine whether AAV could elicit these effects on other cell

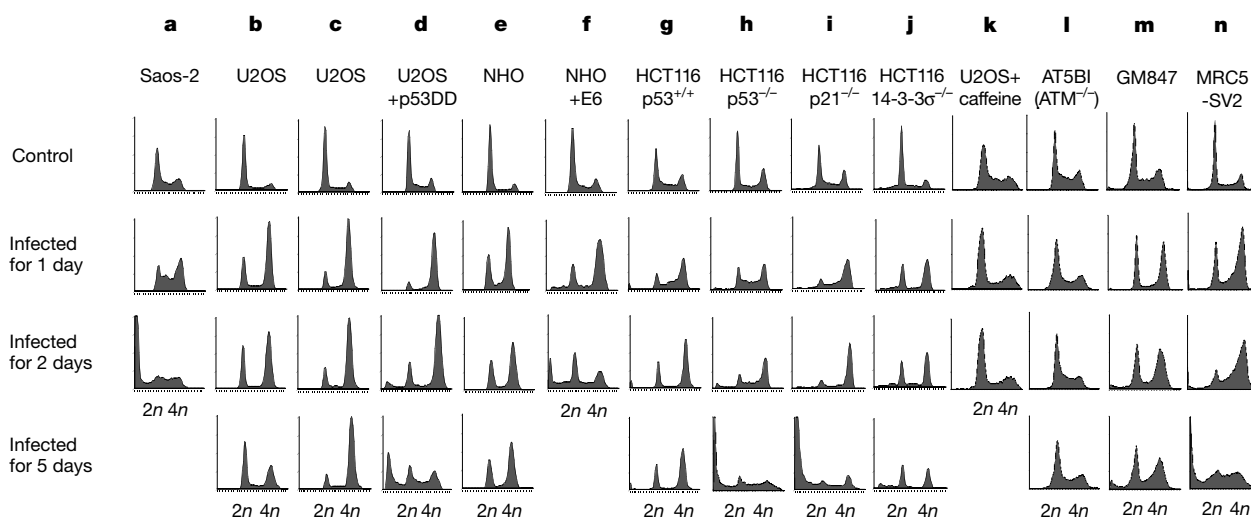


Figure 1 DNA content of cells after AAV infection. Cells were infected with AAV at a multiplicity of 250 (a and b) or 5,000 (c–n). After the indicated times, cells were collected and their DNA content measured by flow cytometry.

types, we infected a pair of isogenic human colon carcinoma cell lines, HCT116p53^{+/+} and HCT116p53^{-/-} (ref. 1), with AAV. Both the HCT116p53^{+/+} and HCT116p53^{-/-} cells arrested at the G2 phase of the cell cycle (Fig. 1g and h). However, the HCT116p53^{-/-} cells underwent cell death after several days (Fig. 1h). In AAV-infected HCT116p53^{+/+} cells, the levels of p53, p21 and 14-3-3 σ proteins increased whereas that of CDC25C decreased (Fig. 2g). Interestingly, the level of CDC25C protein in HCT116p53^{-/-} cells, like that in the U2OS+p53DD cells, was unchanged after AAV infection (Fig. 2h), once again revealing the coupling of CDC25C protein

degradation to p53 activity. To test the relative importance of p21 and 14-3-3 σ , HCT116 cells that lacked either of these proteins were infected with AAV. Cells lacking p21 failed to sustain the G2 arrest and died, while those lacking 14-3-3 σ were still able to sustain the arrest with minimal cell death (Fig. 1i and j), underlining the importance of p21 in the G2 arrest.

The cellular responses to AAV and to DNA damage bear many similarities. In both cases, cells can respond either by establishing an arrest at G2 if p53 is present, or by pausing briefly at G2 before dying, when p53 is absent^{1,2}. We treated U2OS cells with etoposide, a DNA-damaging drug¹⁰, in place of AAV infection. When the levels of p53, p21 and CDC25C proteins were analysed, they were found to change in a manner similar to that caused by AAV (Fig. 2i). We know that, in response to DNA damage, nuclear localization of CDC25C is inhibited, inducing a G2 arrest¹¹. Our results reveal another mechanism for inactivation of CDC25C phosphatase: proteasomal degradation. An early step in DNA damage signalling can be mediated by the ATM protein¹². U2OS cells infected with AAV in the presence of the ATM inhibitor caffeine¹³ failed to arrest at the G2 phase, but continued to proliferate (Fig. 1k). Consistent with this, ATM-null cells (AT5BI, transformed with simian virus 40 (SV40)) were not affected by AAV. The SV40-transformed control cells (GM847 and MRC5-SV2) accumulated in G2 with cell death presumably depending on the level of residual p53 activity (Fig. 1l–n). These findings suggest that AAV affects the cell by inducing an ATM-dependent DNA damage response. Failure to induce this in ATM-null cells renders AAV ineffective in causing G2 arrest or apoptosis.

To test whether the incoming viral proteins were responsible for these effects, we used AAV-like particles without DNA we made using recombinant baculoviruses¹⁴, or empty AAV particles containing Rep, but not AAV DNA. When exposed to ultraviolet irradiation and added to cells, none of these affected cell growth (data not shown). Retroviral-mediated expression of the Rep proteins in cells did not induce the effects we saw with AAV¹⁵. The ultraviolet-inactivated AAV used here is unable to support the synthesis of viral proteins or DNA. As the effect of AAV on cells was not diminished by ultraviolet treatment, we conclude that viral replication or gene expression is not required for these effects. Ultraviolet-induced lesions in DNA were not responsible, as equivalent amounts of ultraviolet-inactivated adenovirus had no effect in parallel experiments (data not shown). Instead the evidence indicates that AAV DNA, which is single stranded with hairpin loops at both ends, can be sensed as abnormal DNA by the cell¹⁶ and trigger a DNA damage response. We tested this by injecting Saos-2, U2OS and U2OS+p53DD cells with an oligonucleotide corresponding to the AAV hairpin and containing no AAV coding sequence. This killed

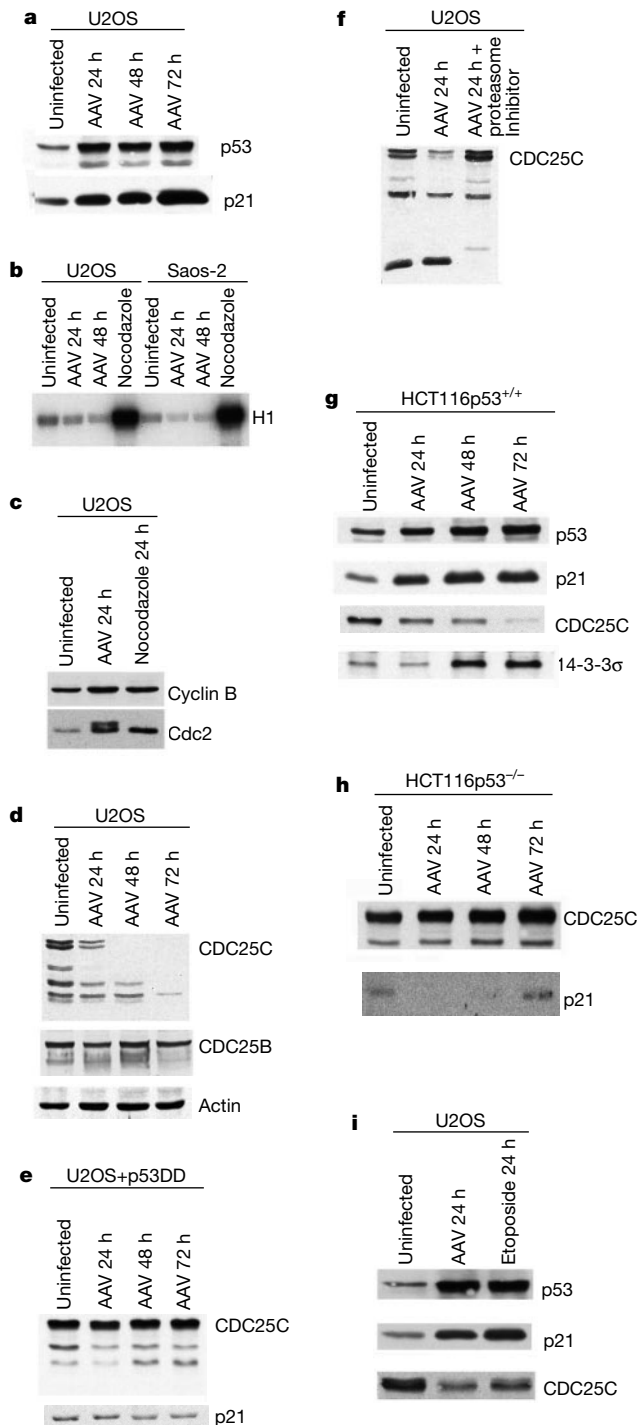


Figure 2 Biochemical analyses of proteins in AAV-infected cells. **a, c–i**, Western blotting analyses of various proteins in the indicated cell types. **b**, Cyclin B–cdc2 kinase assay of cells with and without AAV infection. H1, histone 1.

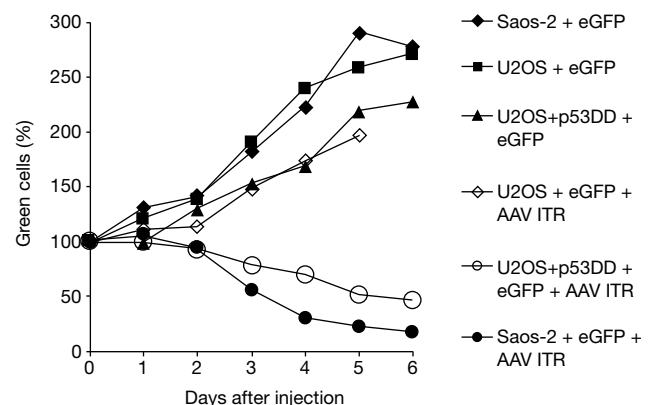


Figure 3 Cell viability following co-injection of AAV hairpin DNA (the inverted terminal repeat, ITR) and plasmids encoding green fluorescence protein (GFP) into different cell types.

Table 1 Effects of AAV on growth of tumours in athymic mice

Effect of AAV on tumour formation			
Cells	Treatment	Tumour incidence	Tumour incidence (%)
HCT116p53 ^{-/-}	PBS	12 of 12	100
	AAV	2 of 12	16.6
HCT116p53 ^{+/+}	PBS	10 of 10	100
	AAV	8 of 10	80
Effect of AAV on established tumours			
Tumours	Treatment	Number of tumours	Relative tumour volume (%)
HCT116 p53 ^{-/-}	PBS	10	100
	AAV	10	34–73
HT29	PBS	10	100
	AAV	4	19–34
		6	0

the Saos-2 and U2OS+p53DD cells but not the U2OS cells (Fig. 3), supporting the conclusion that it is indeed the DNA of AAV that triggers a DNA damage response in the cells.

We tested whether this mechanism may be used to prevent the establishment of tumours. HCT116p53^{-/-} or HCT116p53^{+/+} cells were injected under the skin of nude mice, followed by injection of AAV, or PBS as a control, 2 days later. All of the control HCT116p53^{-/-} injections gave rise to tumours. Tumour incidence was reduced to 17% when AAV was used. HCT116p53^{+/+} cells were more resistant to AAV, as 80% of injections still produced tumours when AAV was present (Table 1 and see Supplementary Information). No non-specific toxicity was observed. This result is consistent with the observation¹⁷ that defective AAV particles that contained sub-genomic-length DNA with terminal hairpins protected hamsters against adenovirus-induced tumours. DNA fragments extracted from these defective particles were also able to prevent tumorigenesis. We then asked whether AAV is able to prevent the growth of pre-existing tumours lacking p53. Injection of AAV into established HCT116p53^{-/-} tumours inhibited their growth, resulting in tumours smaller than those injected with PBS. AAV injected into tumours derived from HT29, another p53-null colon carcinoma cell line, caused complete regression of 60% of the tumours and suppressed the growth of the remainder (Table 1 and see Supplementary Information).

How human tumours that lack p53 activity die in response to therapy remains puzzling¹⁸. Activation of p53 is often considered to lead to cell death, notably in lymphocytes¹⁹. However, there are cell types and situations where the presence of p53 prevents cell death by DNA damage^{1,2,19}. As irradiation and genotoxic drugs cause physical damage to cellular DNA, it was proposed that when damaged p53-null cells attempt to divide, they undergo mitotic catastrophe². In the experiments described above, we induced a DNA damage response using AAV instead of damaging the DNA of the cell, and observed that cells which lack p53 activity arrest transiently and die, suggesting that they possess a checkpoint that triggers apoptosis if they fail to sustain a G2 arrest in the face of DNA damage signalling. The absence of damaged cellular DNA may also explain why the ATM^{-/-} cells did not establish a delayed G2 arrest, as reported for certain ATM^{-/-} cells following irradiation²⁰.

Whereas further work will elucidate what kind of DNA damage AAV DNA is mimicking and details of which pathway is being triggered, the immediate importance of these findings is the introduction of the principle of using viruses to efficiently deliver DNA with unusual or modified structures into cells to eliminate those cells that lack p53 activity. □

Methods

Cell culture and inactivation of p53 *in vivo*

U2OS and Saos-2 cells were cultured in DMEM supplemented with 10% fetal calf serum. NHO cells were purchased from Clonetics and were cultured in Osteoblast Growth

Medium (Clonetics) supplemented with 10% fetal calf serum and ascorbic acid. DNA encoding the p53DD protein was cloned into the retroviral vector pBabepuro. To inhibit ATM activity, cells were treated with 2 mM caffeine. Annexin V analysis was performed according to the instructions of the manufacturer (Boehringer).

Inactivation of AAV and infection of cells

AAV prepared as described²¹ was diluted in 0.5 ml PBS in a small plastic dish and exposed to 2,400 J m⁻² of ultraviolet irradiation (wavelength 254 nm) from a Stratalinker. The inactivated viruses were further diluted in 2.5 ml DMEM (10% FCS) before adding them to cells. After 3 h, fresh medium was added to make up 10 ml.

Flow cytometry

Cells were gathered and fixed in 70% ethanol. Fixed cells were centrifuged, resuspended and incubated in 100 µg ml⁻¹ RNase in PBS at 37 °C. After 30 min, propidium iodide was added to a concentration of 100 µg ml⁻¹. DNA content was measured with a flow cytometer.

Western blotting and cyclin B-cdc2 kinase assay

Cells were washed twice with PBS and collected with a rubber-covered rod. After centrifugation cell pellets were resuspended in two volumes of Reporter Lysis Buffer (Promega) supplemented with a cocktail of protease inhibitors (Calbiochem). After incubation on ice for 30 min with occasional swirling, the samples were centrifuged at 16,100 g for 10 min. Supernatants were collected and protein concentrations measured by the Bradford assay (BioRad). Protein samples of 30 µg were resolved on SDS-polyacrylamide gels, transferred to nylon membranes (Hybond) and analysed with antibodies against p53, p21, CDC25C, CDC25B, actin, cyclin B and cdc2 (all from Santa Cruz). Cyclin B-cdc2 kinase assays were performed as described²².

Injection of cells

DNA used in these experiments were first passed through a 0.2-µm filter. pCieGFP is a plasmid that contains a CMV promoter which controls the expression of the green fluorescent protein gene. The AAV hairpin oligonucleotide was synthesized (Microsynth) based on the sequence of the AAV-2 inverted terminal repeat (nucleotide positions 1–145). DNAs (either pCieGFP at 400 µg ml⁻¹, or pCieGFP at 200 µg ml⁻¹ with hairpin DNA at 200 µg ml⁻¹) were injected into cells with an Eppendorf Micromanipulator. Green cells were visible 4 h after injection and were then counted on successive days.

Received 30 March; accepted 11 June 2001.

- Bunz, F. *et al.* Requirement for p53 and p21 to sustain G2 arrest after DNA damage. *Science* **282**, 1497–1501 (1998).
- Chan, T. A., Hermeking, H., Lengauer, C., Kinzler, K. W. & Vogelstein, B. 14-3-3σ is required to prevent mitotic catastrophe after DNA damage. *Nature* **401**, 616–620 (1999).
- Srivastava, A., Lusby, E. W. & Berns, K. I. Nucleotide sequence and organisation of the adeno-associated virus 2 genome. *J. Virol.* **45**, 555–564 (1983).
- Berns, K. I. & Linden, R. M. The cryptic life cycle of adeno-associated virus. *Bioessays* **17**, 237–245 (1995).
- Bowman, T. *et al.* Tissue specific inactivation of p53 tumour suppression in the mouse. *Genes Dev.* **10**, 826–835 (1996).
- Xiong, Y. *et al.* p21 is a universal inhibitor of cyclin-dependent kinases. *Nature* **366**, 701–704 (1993).
- Nurse, P. Universal control mechanism regulating cell cycle timing in M-phase. *Nature* **344**, 503–508 (1990).
- Strausfeld, U. *et al.* Activation of p34cdc2 protein kinase by microinjection of human CDC25C into mammalian cells. Requirement for prior phosphorylation of cdc25C by p34cdc2 on sites phosphorylated at mitosis. *J. Biol. Chem.* **269**, 5989–6000 (1994).
- Rock, K. *et al.* Inhibitors of the proteasome block the degradation of most cell proteins and the generation of peptides presented on MHC class I molecules. *Cell* **78**, 761–771 (1994).
- Wozniak, A. J. & Ross, W. E. DNA damage as a basis for 4'-demethylpiperidophyllotoxin-9-(4,6-O-ethylidene-beta-D-glucopyranoside) (etoposide) cytotoxicity. *Cancer Res.* **43**, 120–124 (1983).
- Peng, C. Y. *et al.* Mitotic and G2 checkpoint control regulation of 14-3-3 protein binding by phosphorylation of CDC25C on serine-216. *Science* **277**, 1501–1505 (1997).
- Rotman, G. & Shiloh, Y. ATM: a mediator of multiple responses to genotoxic stress. *Oncogene* **18**, 6135–6144 (1999).
- Blasina, A., Price, D. D., Turenne, G. A. & Gowen, C. H. Caffeine inhibits the checkpoint kinase ATM. *Curr. Biol.* **9**, 1135–1138 (1999).
- Ruffing, M., Zentgraf, H. & Kleinschmidt, J. A. Assembly of viruslike particles by recombinant structural proteins of adeno-associated virus type 2 in insect cells. *J. Virol.* **66**, 6922–6930 (1992).
- Saudan, P., Vlach, J. & Beard, P. Inhibition of S-phase progression by adeno-associated virus is mediated by hypophosphorylated pRb. *EMBO J.* **19**, 4351–4361 (2000).
- Yang, D., van Boom, S. S., Reedijk, J., Farrell, N. & Wang, A. H. A novel DNA structure induced by the anticancer bisplatinum compound crosslinked to a GpC site in DNA. *Nature Struct. Biol.* **2**, 577–578 (1995).
- De La Maza, L. M. & Carter, B. J. Inhibition of adenovirus oncogenicity in hamsters by adeno-associated virus DNA. *J. Natl Cancer Inst.* **67**, 1323–1326 (1981).
- Finkel, E. Does cancer therapy trigger cell suicide? *Science* **286**, 2256–2258 (1999).
- Brown, J. M. & Wouters, B. G. Apoptosis, p53, and tumour cell sensitivity to anticancer agents. *Cancer Res.* **59**, 1391–1399 (1999).
- Beamish, H. & Lavin, M. F. Radiosensitivity in ataxia-telangiectasia: anomalies in radiation-induced cell cycle delay. *Int. J. Radiat. Biol.* **65**, 175–184 (1994).
- Ogston, P., Raj, K. & Beard, P. Productive replication of adeno-associated virus can occur in human papillomavirus type 16 (HPV16) episome-containing keratinocytes and is augmented by the HPV16 E2 protein. *J. Virol.* **74**, 3494–3504 (2000).

22. Fournier, N. *et al.* Expression of human papillomavirus 16 E2 protein in *Schizosaccharomyces pombe* delays the initiation of mitosis. *Oncogene* **18**, 4015–4021 (1999).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank N. Paduawat and B. Bentele for technical assistance; M. Chapman and X. Qing for the gift of AAV-2; B. Vogelstein for the HCT116 cell line and its derivatives; R. Iggo for antibodies and the Saos-2 and U2OS cell lines; M. Oren for p53DD vectors; A. M. R. Taylor for the ATM-null cells; K. Alevizopoulos for retroviral vectors; J. Kleinschmidt for recombinant baculovirus; C. Wenz for assistance in insect cell culture; and S. Gasser, A.-L. Ducrest, M. Höss, B. Hirt, J. Jiricny, R. Iggo and V. Simanis for discussions. B. Sordat and F. Hoffmann helped with the animal experiments, and P. Reichenbach and J. Wyniger with the micro-injection. This work was supported by Cancer Research Switzerland and the Swiss National Science Foundation.

Correspondence and requests for materials should be addressed to P.B. (e-mail: peter.beard@isrec.unil.ch).

Rotational movement during cyclic nucleotide-gated channel opening

J. P. Johnson Jr & William N. Zagotta

Howard Hughes Medical Institute & Department of Physiology and Biophysics, Box 357290, University of Washington School of Medicine, Seattle, Washington 98195, USA

Cyclic nucleotide-gated (CNG) channels are crucial components of visual, olfactory and gustatory signalling pathways. They open in response to direct binding of intracellular cyclic nucleotides and thus contribute to cellular control of both the membrane potential and intracellular Ca^{2+} levels¹. Cytosolic Ni^{2+} potentiates the rod channel (CNG1) response to cyclic nucleotides^{2–4} and inhibits the olfactory channel (CNG2) response⁵. Modulation is due to coordination of Ni^{2+} by channel-specific histidines in the C-linker, between the S6 transmembrane segment and the cyclic nucleotide-binding domain. Here we report, using a histidine scan of the initial C-linker of the CNG1 channel, stripes of sites producing Ni^{2+} potentiation or Ni^{2+} inhibition, separated by 50° on an α -helix. These results suggest a model for channel gating where rotation of the post-S6 region around the channel's central axis realigns the Ni^{2+} -coordinating residues of multiple subunits. This rotation probably initiates movement of the S6 and pore opening.

The mechanism of Ni^{2+} modulation is simple: Ni^{2+} binding is state dependent. Ni^{2+} is an intermediate Lewis acid and prefers association with intermediate Lewis bases, such as the imidazole of histidine⁶. In CNG1 channels, naturally bearing H420, the open channel has a higher affinity (>100 times) for Ni^{2+} than the closed channel, and Ni^{2+} promotes channel opening. In CNG2 channels, naturally bearing a histidine at the 417 equivalent position, the closed channel has higher affinity for Ni^{2+} than the open channel, and Ni^{2+} promotes channel closing (Fig. 1a)^{2,5}. Ni^{2+} affects the opening transition, but not the intrinsic ligand binding, as Ni^{2+} modulation is seen even in the presence of saturating concentrations of partial agonists where binding sites are maximally occupied^{2,3,7}. Individual histidines bind Ni^{2+} poorly (~1 mM), but multiple histidines, especially in a square planar geometry, can coordinate Ni^{2+} with sub-micromolar affinity⁸. Ni^{2+} coordination in CNG1 channels occurs between two adjacent channel subunits^{9,10}.

Predictions of the secondary structure¹¹ suggest that the S6 and the C-linker of CNG1 channels are α -helical in nature until position 420. The KcsA K^{+} channel has been crystallized¹², providing a

structural model for the pores of other P-loop-containing cation channels, including CNG channels. To create a working model, we threaded the aligned amino-acid sequence of the S6 and P-loop of the CNG1 channel through the KcsA structure and extended the α -helix of the CNG1 channel S6 to position 420 (Fig. 1b). Residue 417 (Fig. 1b, red) points in the direction of the central axis, which allows Ni^{2+} coordination, while residue 420 (Fig. 1b, green) points away from the central axis. Rotation of the helix on channel opening could change the orientation of the residues such that residue 420 would point towards, and 417 away, from the central axis, thereby explaining state-dependent histidine coordination of Ni^{2+} .

This model predicts that other positions near 417 and 420 in the post-S6 region might also be able to coordinate Ni^{2+} . We completed a histidine scan between positions 403 and 420 in a Ni^{2+} -insensitive CNG1 channel, H420Q (Fig. 2c right)^{2,5}. Each residue between position 403 and 420 was individually replaced with histidine, and cGMP dose-response curves were measured in the absence and presence of Ni^{2+} .

We found three Ni^{2+} phenotypes: Ni^{2+} -inhibited, Ni^{2+} -potentiated and Ni^{2+} -insensitive. Ni^{2+} inhibited three channels, Q409H, D413H and Q417H (Fig. 2a). For Q409H and D413H channels, Ni^{2+} caused a decrease in current at sub-saturating cGMP concentrations, causing the cGMP dose-response relationship to shift right (Fig. 2a, right). Q417H currents were completely inhibited by 1 μM Ni^{2+} , even at saturating concentrations of cGMP, which made the measurement of cGMP dose responses in Ni^{2+} impossible⁵. Ni^{2+} potentiated two channels, K416H and H420 (Fig. 2b). Addition of Ni^{2+} to these channels increased the current, causing the cGMP dose-response relationship to shift left (Fig. 2b, right). Ni^{2+} had no effect on 11 out of 18 channel constructs (Fig. 2c). Two mutants (I415H and M419H) failed to produce currents in either the absence or presence of Ni^{2+} .

To quantify the effect of intracellular Ni^{2+} , we fit the cGMP dose-response relationships for all mutants (except D413H) in the absence and presence of Ni^{2+} with a simple allosteric model containing two ligand-binding events and an allosteric opening transition (see Methods). Although the model is oversimplified, it provides a reasonable fit of the data and a straightforward analysis.

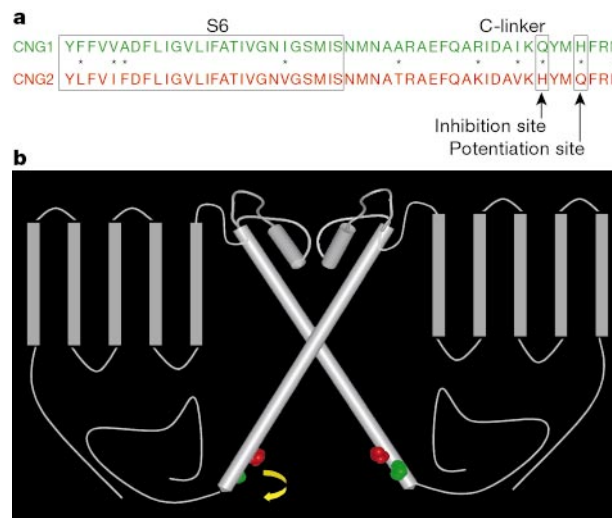


Figure 1 Ni^{2+} -coordinating histidines. **a**, Sequence alignment showing the location of Ni^{2+} -coordinating histidines in CNG1 and CNG2. Stars indicate nonconserved positions. **b**, Depiction of two of the four CNG1 channel subunits. S1–S5 transmembrane domains are represented by vertical grey rectangles. The cylinders representing the pore and S6 helices are the result of aligning the CNG1 sequence with the KcsA channel structure and extending the S6 to position 420, although sequence similarity with KcsA ends near the helical bundle crossing at 399. Residues 417 and 420 are coloured red and green, respectively.

D413H channels produce a shallow dose-response relationship (Hill slope near 1) that cannot be fit with the two-binding-step model, and thus a single-binding-step model was used for D413H. Assuming a fixed value for the intrinsic binding affinity (K), the fits yield a measure of the equilibrium constant of allosteric opening (L). Large L values indicate energetically favoured opening transitions. Channels insensitive to Ni^{2+} had very similar L values in the presence and absence of Ni^{2+} (Fig. 3a). Mutants inhibited by Ni^{2+} , Q409H and Q413H, had lower L values in Ni^{2+} than in its absence. Presumably Q417H would behave similarly if L in the presence of Ni^{2+} could be measured, as is the case when histidine is present in the equivalent position of CNG2 channels³. Ni^{2+} -potentiated mutants, K416H and H420, showed opposite results— L values in Ni^{2+} were higher than in its absence.

The ratio of L values in the absence and presence of Ni^{2+} were used to calculate the Ni^{2+} -induced change in free energy of the allosteric transition, $\Delta\Delta G_{\text{Ni}}$ (Fig. 3b; see Methods). Ni^{2+} -insensitive mutants had $\Delta\Delta G_{\text{Ni}}$ values near 0. Ni^{2+} -inhibited mutants had positive $\Delta\Delta G_{\text{Ni}}$ values, as Ni^{2+} favoured the closed state; Ni^{2+} -potentiated mutants had negative $\Delta\Delta G_{\text{Ni}}$ values, as Ni^{2+} favoured the open state. Ni^{2+} sensitivity was periodic, with inhibiting sites every four positions and potentiating sites every four positions.

All of the amino acids studied are well outside of the cyclic nucleotide-binding domain and were not expected to change the intrinsic ligand-binding affinity. However, it was possible that the changes in cGMP dose-responses were due to changes in intrinsic ligand binding. To test this, we measured current evoked by saturating concentrations of the full agonist cGMP and the partial agonists cAMP and cIMP with and without Ni^{2+} (Fig. 4a). Saturating concentrations of any cyclic nucleotide should maximally occupy the cyclic nucleotide-binding sites; therefore, changes in the intrinsic ligand affinity would not affect L values measured in saturating ligand. We tested all Ni^{2+} -sensitive mutants from our initial experiments plus the control channel (H420Q). Assuming that the mutations do not have cyclic nucleotide-specific effects on

the L value (that is, mutant and Ni^{2+} effects on $\Delta\Delta G_{\text{Ni}}$ are equivalent for all cyclic nucleotides), we determined L values for the mutants independent of a ligand-binding model (Fig. 4b; see Methods). As before, channels inhibited by Ni^{2+} had smaller L values in the presence of Ni^{2+} , whereas channels potentiated by Ni^{2+} had larger L values in Ni^{2+} . Ni^{2+} had no effect on the L value of the H420Q channel. $\Delta\Delta G_{\text{Ni}}$ values calculated from saturating cyclic nucleotide data were similar to $\Delta\Delta G_{\text{Ni}}$ values calculated from cGMP dose-responses (Figs 4c and 3b, respectively). Small differences in the $\Delta\Delta G_{\text{Ni}}$ values calculated by these two methods may be due to the simplified nature of the gating model.

Ni^{2+} -inhibited positions were found every fourth amino acid (409, 413, 417). Potentiating positions (416, 420) were also four amino acids apart. Mapping the results of our scan on the homology model (Fig. 1b) generates Fig. 5a. Residue S399 (Fig. 5a, yellow) is at a position equivalent to the bottom of the helical bundle of KcsA channels. A S399C mutation results in intersubunit disulphide bonds in closed, but not open, CNG1 channels, suggesting that 399 residues are closer together in closed than in open channels¹³. The model in Fig. 5a assumes that the S6 helix continues unbroken through the initial C-linker. Inhibiting (red) and potentiating (green) positions fell along two vertical stripes on the helix. The stripe of inhibiting residues is offset 50° from the potentiating stripe. Clockwise rotation of the stripes relative to the central axis (as viewed extracellularly) on opening would move potentiating residues towards the place occupied by inhibiting residues in closed channels.

Although the model in Fig. 5a is intuitive, there are potential complications. The distances between coordinating residues are large (20–40 Å), too large to facilitate coordination of Ni^{2+} (~10 Å)^{8,14}. Another problem is the orientation of residues on the helix. This model predicts reactive stripes that are parallel to the helical axis if the coordinating residues are to be in the same orientation relative to the central axis of the channel. This is not the observed pattern of every four residues, where reactive stripes

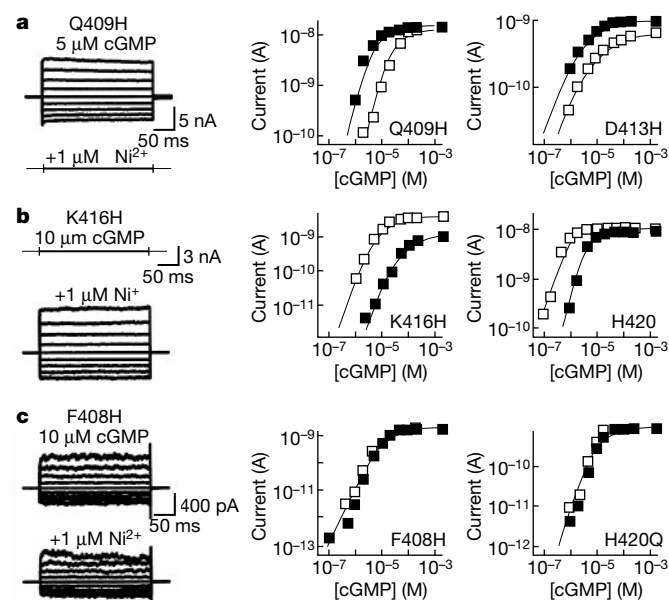


Figure 2 Ni^{2+} -histidine modulation. **a**, Q409H currents (left) in 5 μM cytosolic cGMP without and with 1 μM Ni^{2+} . cGMP dose-response curves (right) are shown without (filled squares) and with (open squares) 1 μM Ni^{2+} (Q409H, D413H). **b**, K416H currents (left) as in **a** in 10 μM cytosolic cGMP without and with 1 μM Ni^{2+} . cGMP dose-response curves are shown as in **a** (K416H, H420). **c**, F408H currents (left) as in **a** in 10 μM cGMP without and with 1 μM Ni^{2+} . cGMP dose-response curves (right) are shown as in **a** for F408H and the control channel, H420Q.

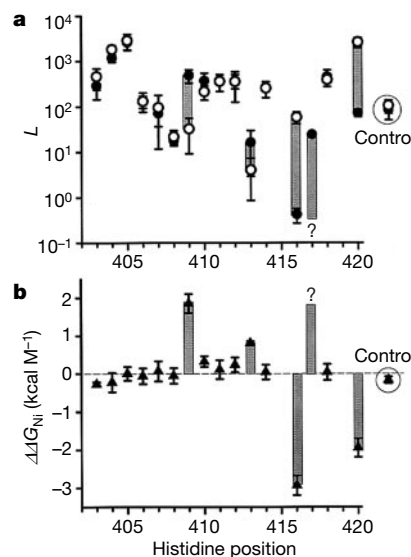


Figure 3 Energetic effects of Ni^{2+} . **a**, Equilibrium constants of the allosteric opening transition (L) without (filled circles) and with (open circles) 1 μM Ni^{2+} . L values are shown versus histidine position. L values for the control H420Q are circled on the right. Grey bars highlight the fold change in L for Ni^{2+} -sensitive mutants. **b**, $\Delta\Delta G_{\text{Ni}}$ values for the opening transition were calculated from L values in **a**. Grey bars highlight $\Delta\Delta G_{\text{Ni}}$ values for Ni^{2+} -sensitive mutants. $\Delta\Delta G_{\text{Ni}}$ for the control H420Q channel is circled on the right. One micromolar of Ni^{2+} abolished Q417H currents, thus neither L in Ni^{2+} nor $\Delta\Delta G_{\text{Ni}}$ could be determined (question marks), but a large, positive $\Delta\Delta G_{\text{Ni}}$ is indicated. Error bars indicate s.e.m.

cross the helical faces 35° off the helical axis. The observed pattern of Ni^{2+} reactivity is exactly as would be expected if these residues are at a right-handed, helical bundle crossing—analogue to the helical bundle crossing of KcsA¹². A possibility that would overcome both problems of distance and exposure angle would be if the helix were kinked between positions 399 and 409. Inserting a kink by rotating the ϕ - and ψ -angles of the peptide backbone at position 403 results in a 'kinky' model (Fig. 5b). The kink re-orientates inhibiting

residues to directly face the central axis and gives them a shorter, more consistent intersubunit distance. Kinking also results in a second helical bundle crossing such that the pattern of diagonal stripes across the helical axes is explained.

Our data suggest that channel opening involves rotation of the initial C-linker around the central axis without translating radially (Fig. 5b, c). Reactive amino acids would effectively rotate clockwise relative to the central axis during opening. This movement would

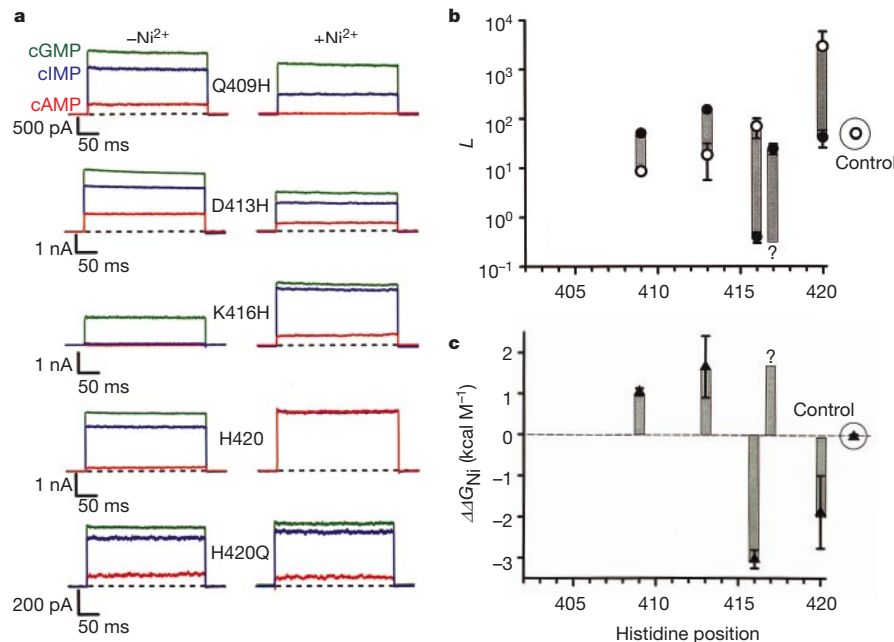


Figure 4 L values calculated from saturating cyclic nucleotide concentrations. **a**, Currents induced by saturating cGMP (2 mM, green), cIMP (16 mM, blue), and cAMP (16 mM, red) at +60 mV are shown for Ni^{2+} -sensitive channels without and with $1 \mu\text{M}$ Ni^{2+} . **b**, L values calculated from saturating ligand currents without (filled circles) and with (open

circles) $1 \mu\text{M}$ Ni^{2+} . **c**, $\Delta\Delta G_{\text{Ni}}$ values for the opening transition are plotted versus histidine position for Ni^{2+} -sensitive mutants. H420Q (control) values are circled on the right in **b** and **c**. One micromolar of Ni^{2+} abolished Q417H currents, thus neither L in Ni^{2+} nor $\Delta\Delta G_{\text{Ni}}$ could be determined for this mutant (question marks). Error bars indicate s.e.m.

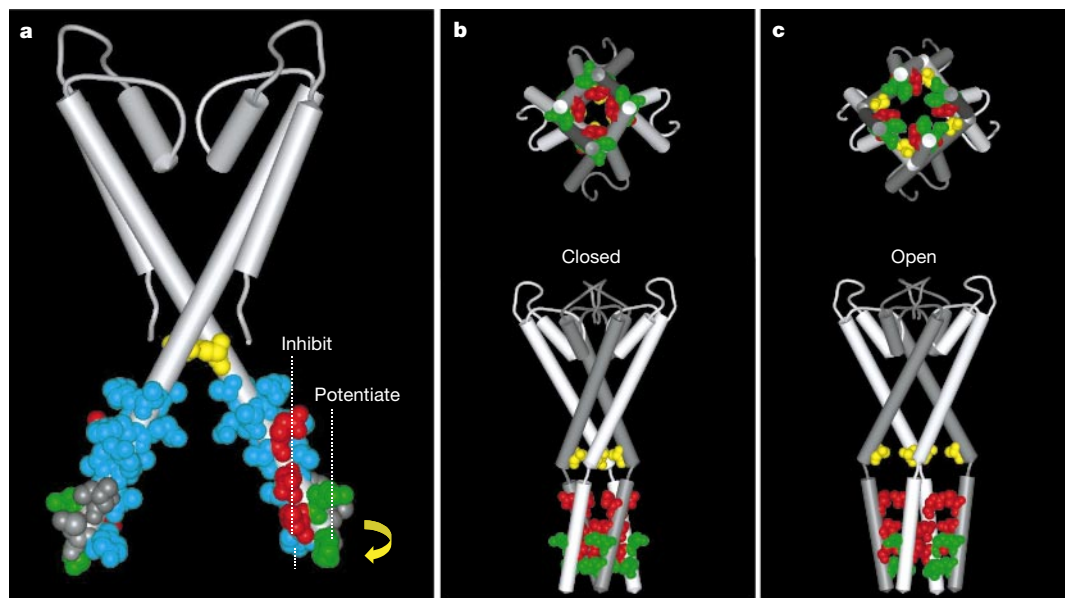


Figure 5 Models showing how Ni^{2+} coordination could change CNG1 gating. **a**, Ni^{2+} -sensitivity results are shown mapped on the model of Fig. 1a. Histidine substitution at the red positions produced Ni^{2+} -inhibited channels. Histidine substitution at the green positions produced Ni^{2+} -potentiated channels. Position 399 is yellow. Blue residues

produced Ni^{2+} -insensitive channels. Mutating grey residues yielded non-functional channels. The kinky model of channel gating is shown in the closed (**b**) and open (**c**) states.

allow the inhibiting positions to coordinate Ni^{2+} when the channel is closed and the potentiating positions to coordinate Ni^{2+} when the channel is open. It would also increase radial separation of 399 residues and widen the S6 bundle crossing—movements that are associated with channel opening¹³. The movement of the C-linker is likely to control, or at least report, movement of the associated S6 transmembrane domain. Neither Ni^{2+} inhibition nor potentiation was voltage dependent and the Ni^{2+} concentration used does not cause appreciable block of the channel. We therefore expect that Ni^{2+} -binding residues are more cytoplasmic than the permeation pathway and that permeant ions flow through 'windows' below the S6 bundle crossing in the model.

Our data provide indications of the nature of movement in the CNG channel C-linker. The C-linker is critical to the coupling of ligand binding to channel opening^{2,5,15–19}. We show the C-linker bundle as a four-fold symmetrical structure, but our data do not eliminate the possibility that the bundles actually associate as a dimer of dimers. It is known that histidine- Ni^{2+} coordination by H420 is mediated by adjacent but not opposite subunits¹⁰, and there is evidence that the channel gates as a dimer of dimers²⁰.

The kinky model (Fig. 5b, 5c) is reminiscent of a model recently proposed for KcsA based on EPR data, where channel opening involves a translation and rotation of TM2 (refs 21, 22). More recently KcsA was suggested to be kinked just after TM2, causing a second, intracellular bundle crossing²³. Our data suggest that CNG1 channels also contain a second bundle crossing, which is cytoplasmic to S6, in the C-linker. We propose that a rotational movement at this bundle crossing is associated with channel opening. The CNG channel C-linkers probably act as lever arms, coupling conformational changes of the ligand-binding domains to pore opening. □

Methods

Mutagenesis

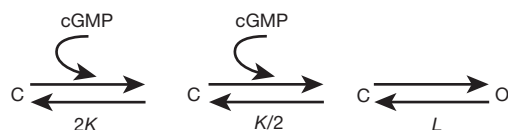
All channel constructs were made from a modified version of CNG1 in which all endogenous cysteine codons have been mutated²⁴, and H420 was replaced with glutamine. Mutagenesis was accomplished by standard methods². The PCR-amplified regions of all mutants were sequenced completely.

Oocyte expression and electrophysiology

CNG1 complementary DNAs in the pGEMHE plasmid were linearized with *NheI* restriction enzyme, and RNA was synthesized *in vitro* and injected into defolliculated *Xenopus laevis* oocytes as described²⁵. Recordings were made in the excised, inside-out patch clamp configuration using an Axopatch 200A patch-clamp amplifier (Axon Instruments) and Pulse acquisition software (HEKA Elektronik)²⁶. Patch pipettes were pulled from borosilicate glass and had resistances of 0.25–1 MΩ after fire polishing. We analysed data using a combination of Excel (Microsoft), Igor Pro (Wavemetrics) and SigmaPlot (Jandel Scientific). The extracellular pipette solution for all recordings was (in mM): 130 NaCl, 0.5 niflumic acid, 0.2 EDTA, 3 HEPES, pH 7.2. The control intracellular solution was (in mM): 130 NaCl, 0.2 EDTA, 3 HEPES, pH 7.2. Cyclic nucleotides were diluted from 16 mM stock solutions to the desired concentration in this solution. Experiments were begun 20–30 min after excision, after current run up was complete. For Ni^{2+} experimental solutions, EDTA was omitted and 1 μM NiCl_2 was included. Patches were initially exposed to 10 μM NiCl_2 for 5 min to ensure complete Ni^{2+} modulation, and subsequent measurements were made in 1 μM NiCl_2 . All records were corrected for leak by subtraction of the current recorded from the same patch under the same ionic conditions but in the absence of cyclic nucleotides. The voltage-clamp protocol was always the same: 300-ms test pulses between +80 and –80 mV were applied in 20-mV increments at 1-s intervals from a holding potential of 0 mV. All of the measurements were made during the voltage step to +60 mV.

Secondary structure prediction and analysis

Secondary structure was predicted using the nnpredict program (<http://www.cmpharm.ucsf.edu/~nmi/nnpredict-instrucs.html>)¹¹. The dose-response data (except D413H) were fit by the linear model in scheme 1.

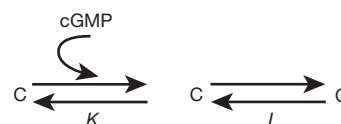


This model includes two independent cyclic nucleotide-binding steps followed by a concerted allosteric opening event. K is the association equilibrium constant for binding of cGMP; L is the equilibrium rate constant of the allosteric transition. The assumption of two binding steps is required to fit the slope of the dose-response data. The dose-response relationships for cGMP were fit by equation (1).

$$I = I_{\max} \frac{[\text{cGMP}]^2 L K^2}{1 + 2[\text{cGMP}]K + [\text{cGMP}]^2 K^2 + [\text{cGMP}]^2 K^2 L} \quad (1)$$

where I is the current and $I_{\max}$ is the current when the open probability is 1. K was determined by fitting multiple mutants simultaneously and K was then held fixed at a value of 16,000 M⁻¹; $I_{\max}$ and L were allowed to vary in the fit.

Equation (1) could not fit the D413H data as the cGMP dose-response relationship for that mutant was shallower than that of the other channel constructs. D413H channel data were therefore fit with the model in scheme (2).



This model includes a single cyclic nucleotide-binding step followed by a concerted allosteric opening event and is described by equation (2).

$$I = I_{\max} \frac{[\text{cGMP}]KL}{1 + [\text{cGMP}]K + [\text{cGMP}]KL} \quad (2)$$

The data at saturating cyclic nucleotide concentrations were fit by equation (3).

$$I_{\text{CNMP}} = I_{\max} \frac{L_{\text{cGMP}} \left(\frac{L_{\text{CNMP}}}{L_{\text{cGMP}}} \right)}{L_{\text{cGMP}} \left(\frac{L_{\text{CNMP}}}{L_{\text{cGMP}}} \right) + 1} \quad (3)$$

Where I_{CNMP} is I_{cGMP} , I_{cAMP} or I_{cIMP} the experimentally determined macroscopic current with a saturating concentration of cyclic nucleotide (2.0 mM for cGMP and 16 mM for cAMP and cIMP); $I_{\max}$ is the current if the channel open probability was 1. L_{CNMP} is L_{cGMP} , L_{cAMP} or L_{cIMP} . We assumed that the nucleotide selectivity and hence the ratios of $L_{\text{cAMP}}:L_{\text{cGMP}}$ and $L_{\text{cIMP}}:L_{\text{cGMP}}$ were the same for all mutants tested. This assumption is reasonable as none of these mutations are near the cyclic nucleotide-binding domain. Simultaneous fitting of data from several mutants yielded ratios for $L_{\text{cIMP}}:L_{\text{cGMP}}$ (0.06) and $L_{\text{cAMP}}:L_{\text{cGMP}}$ (0.004) that were then held constant for fitting L . Using these ratios and the measured currents for I_{cGMP} , I_{cAMP} and I_{cIMP} we then fit currents to equation (3) using a least squares algorithm.

The change in free energy of allosteric opening on addition of Ni^{2+} , $\Delta\Delta G_{\text{Ni}}$, was calculated from the experimentally determined L values using equation (4).

$$\Delta\Delta G_{\text{Ni}} = -RT \ln \frac{L_{\text{Ni}}}{L} \quad (4)$$

where R is the gas constant and T is absolute temperature. L is measured in the absence of Ni^{2+} , and L_{Ni} is measured in the presence of Ni^{2+} .

Structural model creation

The structural diagrams were made using a combination of WebLab ViewerPro v. 3.7 (<http://www.msi.com>), Swiss-Pro and Swiss-PdbViewer (<http://www.expasy.ch>).

Received 11 May; accepted 2 July 2001.

1. Yau, K. W. & Baylor, D. A. Cyclic GMP-activated conductance of retinal photoreceptor cells. *Annu. Rev. Neurosci.* **12**, 289–327 (1989).
2. Gordon, S. E. & Zagotta, W. N. A histidine residue associated with the gate of the cyclic nucleotide-activated channels in rod photoreceptors. *Neuron* **14**, 177–183 (1995).
3. Ildefonse, M., Crouzet, S. & Bennett, N. Gating of retinal rod cation channel by different nucleotides: comparative study of unitary currents. *J. Membr. Biol.* **130**, 91–104 (1992).
4. Karpen, J. W., Brown, R. L., Stryer, L. & Baylor, D. A. Interactions between divalent cations and the gating machinery of cyclic GMP-activated channels in salamander retinal rods. *J. Gen. Physiol.* **101**, 1–25 (1993).
5. Gordon, S. E. & Zagotta, W. N. Localization of regions affecting an allosteric transition in cyclic nucleotide-activated channels. *Neuron* **14**, 857–864 (1995).
6. Cowan, J. A. *Inorganic Biochemistry: An Introduction* (Wiley-VCH, New York, 1997).
7. Sunderman, E. R. & Zagotta, W. N. Mechanism of allosteric modulation of rod cyclic nucleotide-gated channels. *J. Gen. Physiol.* **113**, 601–620 (1999).
8. Maroney, M. J. Structure/function relationships in nickel metallobiochemistry. *Curr. Opin. Chem. Biol.* **3**, 188–199 (1999).
9. Gordon, S. E. & Zagotta, W. N. Subunit interactions in coordination of Ni^{2+} in cyclic nucleotide-gated channels. *Proc. Natl Acad. Sci. USA* **92**, 10222–10226 (1995).
10. Shammatt, I. M. & Gordon, S. E. Stoichiometry and arrangement of subunits in rod cyclic nucleotide-gated channels. *Neuron* **23**, 809–819 (1999).
11. Kneller, D. G., Cohen, F. E. & Langridge, R. Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.* **214**, 171–182 (1990).

12. Doyle, D. A. *et al.* The structure of the potassium channel: molecular basis of K⁺ conduction and selectivity. *Science* **280**, 69–77 (1998).
13. Flynn, G. E. & Zagotta, W. N. Conformational changes in S6 coupled to opening of cyclic nucleotide-gated channels. *Neuron* **30**, 689–698 (2001).
14. Ermler, U., Grabarse, W., Shima, S., Goubeaud, M. & Thauer, R. K. Active sites of transition-metal enzymes with a focus on nickel. *Curr. Opin. Struct. Biol.* **8**, 749–758 (1998).
15. Brown, R. L., Snow, S. D. & Haley, T. L. Movement of gating machinery during the activation of rod cyclic nucleotide-gated channels. *Biophys. J.* **75**, 825–833 (1998).
16. Broillet, M. C. A single intracellular cysteine residue is responsible for the activation of the olfactory cyclic nucleotide-gated channel by NO. *J. Biol. Chem.* **275**, 15135–15141 (2000).
17. Gordon, S. E., Oakley, J. C., Varnum, M. D. & Zagotta, W. N. Altered ligand specificity by protonation in the ligand binding domain of cyclic nucleotide-gated channels. *Biochemistry* **35**, 3994–4001 (1996).
18. Zong, X., Zucker, H., Hofmann, F. & Biel, M. Three amino acids in the C-linker are major determinants of gating in cyclic nucleotide-gated channels. *EMBO J.* **17**, 353–362 (1998).
19. Paoletti, P., Young, E. C. & Siegelbaum, S. A. C-linker of cyclic nucleotide-gated channels controls coupling of ligand binding to channel gating. *J. Gen. Physiol.* **113**, 17–34 (1999).
20. Liu, D. T., Tibbs, G. R., Paoletti, P. & Siegelbaum, S. A. Constraining ligand-binding site stoichiometry suggests that a cyclic nucleotide-gated channel is composed of two functional dimers. *Neuron* **21**, 235–248 (1998).
21. Perozo, E., Cortes, D. M. & Cuello, L. G. Structural rearrangements underlying K⁺-channel activation gating. *Science* **285**, 73–78 (1999).
22. Perozo, E., Cortes, D. M. & Cuello, L. G. Three-dimensional architecture and gating mechanism of a K⁺ channel studied by EPR spectroscopy. *Nature Struct. Biol.* **5**, 459–469 (1998).
23. Cortes, D. M., Cuello, L. G. & Perozo, E. Molecular architecture of full-length KcsA. role of cytoplasmic domains in ion permeation and activation gating. *J. Gen. Physiol.* **117**, 165–180 (2001).
24. Matulef, K., Flynn, G. E. & Zagotta, W. N. Molecular rearrangements in the ligand-binding domain of cyclic nucleotide-gated channels. *Neuron* **24**, 443–452 (1999).
25. Zagotta, W. N., Hoshi, T. & Aldrich, R. W. Gating of single Shaker potassium channels in *Drosophila* muscle and in *Xenopus* oocytes injected with Shaker mRNA. *Proc. Natl Acad. Sci. USA* **86**, 7243–7247 (1989).
26. Hamill, O. P., Marty, A., Neher, E., Sakmann, B. & Sigworth, F. J. Improved patch-clamp techniques for high-resolution current recording from cells and cell-free membrane patches. *Pflügers Arch. Eur. J. Physiol.* **391**, 85–100 (1981).

Acknowledgements

We would like to thank K. D. Black, G. Sheridan and especially H. Utsugi for technical assistance. We also thank A. D. Bragg, K. Craven, G. E. Flynn, K. Matulef, M. C. Trudeau and J. Zheng for their insightful discussions and critical reading of the manuscript.

Correspondence and requests for materials should be addressed to W.N.Z.
(e-mail: zagotta@u.washington.edu).

Ablation of XRCC2/3 transforms immunoglobulin V gene conversion into somatic hypermutation

Julian E. Sale*, Daniella M. Calandrini*, Minoru Takata†, Shunichi Takeda‡ & Michael S. Neuberger*

* Medical Research Council Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

† Immunology and Molecular Genetics, Kawasaki Medical School, 577 Matsushima, Kurashiki, Okayama 701-0192, Japan

‡ CREST Research Project, Radiation Genetics, Faculty of Medicine, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan

After gene rearrangement, immunoglobulin V genes are further diversified by either somatic hypermutation or gene conversion¹. Hypermutation (in man and mouse) occurs by the fixation of individual, non-templated nucleotide substitutions. Gene conversion (in chicken) is templated by a set of upstream V pseudogenes. Here we show that if the RAD51 paralogues² XRCC2, XRCC3 or RAD51B are ablated the pattern of diversification of the immunoglobulin V gene in the chicken DT40 B-cell lymphoma line³ exhibits a marked shift from one of gene conversion to one of somatic hypermutation. Non-templated, single-nucleotide substitutions are incorporated at high frequency specifically into the

V domain, largely at G/C and with a marked hotspot preference. These mutant DT40 cell lines provide a tractable model for the genetic dissection of immunoglobulin hypermutation and the results support the idea that gene conversion and somatic hypermutation constitute distinct pathways for processing a common lesion in the immunoglobulin V gene. The marked induction of somatic hypermutation that is achieved by ablating the RAD51 paralogues is probably a consequence of modifying the recombination-mediated repair of such initiating lesions.

A small proportion of the gene-conversion events occurring at the immunoglobulin loci in chicken B cells can lead to the generation of non-functional immunoglobulin genes—most frequently through introduction of frameshift mutations^{4–6}. Thus, the generation of surface immunoglobulin-μ (sIgM)-loss variants in DT40 cells can be used to give an initial indication of immunoglobulin variable region (V) gene-conversion activity. Compared with the parental DT40 line, a mutant that lacked RAD54 showed a considerably diminished proportion of sIgM loss variants (Fig. 1). A fluctuation analysis performed on multiple clones revealed that the ΔRAD54 line generates sIgM-loss variants at a frequency nearly tenfold less than that of parental DT40. In contrast, a ΔRAD52 line generates sIgM-loss variants at a similar frequency to wild-type cells (Fig. 1). These observations support earlier findings concerning gene conversion in ΔRAD54- and ΔRAD52-DT40 cells^{7,8}.

We extended our analysis to DT40 cells that lacked XRCC2 and XRCC3. We expected that these RAD51 paralogues, which (as with RAD52 and RAD54) are thought to have an involvement in the recombination-dependent pathway of DNA damage repair^{2,9–12}, might also be implicated in immunoglobulin V gene conversion. Rather than seeing a diminished abundance of sIgM-loss variants, the ΔXRCC2 and ΔXRCC3 lines showed a much greater accumulation of loss variants than the parental line (Fig. 1). (The contrasting effects of RAD54- and XRCC2/3-deficiency on immunoglobulin V gene conversion suggests that, although these proteins are all implicated in recombination-mediated DNA repair, they probably have different functions, which is consistent with the distinct patterns of sensitivity of the deficient cells to different forms of DNA damage^{2,9–12}.) In the case of ΔXRCC2-DT40, transfection of the human XRCC2 complementary DNA under control of the human β-globin promoter caused the frequency of generation of sIgM loss variants to revert to values close to those of the wild type (Fig. 1).

As deficiency in both XRCC2 and XRCC3 is associated with chromosomal instability^{2,13–15}, we wondered whether the increased frequency of sIgM-loss variants could reflect gross rearrangements or deletions within immunoglobulin loci. However, Southern blot analysis of 24 sIgM⁺ subclones of ΔXRCC3-DT40 did not reveal any loss or alteration of the 6-kilobase (kb) *SalI/BamHI* fragment containing the rearranged light-chain variable region (V_λ) (not shown). To ascertain whether more localized mutations in the V gene could account for the loss of sIgM expression, we cloned and sequenced the rearranged V_λ segments in populations of sIgM⁺ cells that were sorted from wild-type, ΔXRCC2- and ΔXRCC3-DT40 subclones after 1 month of clonal expansion.

Sixty-seven per cent of V_λ sequences cloned from the sIgM⁺ subpopulations that were sorted from multiple wild-type DT40 clones carried mutations: in most cases (73%), these mutations rendered the V_λ sequences obviously non-functional (Fig. 2). (Presumably, most of the remaining sIgM⁺ cells carried inactivating mutations either in the heavy-chain variable region (V_H) or outside the sequenced region of V_λ.) As detailed below (see Methods) the mutations could be classified as being attributable to gene conversion templated by an upstream V_λ pseudogene, to non-templated point mutation, or as falling into an ambiguous category. Most (67%) of the inactivating mutations were due to gene conversion although some (15%) were stop codons generated by non-templated point mutations, demonstrating that the low frequency of

point mutations seen in this study (and in others^{4,5}) in DT40 cells is not an artefact of polymerase chain reaction (PCR) but rather reveals that a low frequency of point mutation does indeed accompany gene conversion in wild-type DT40.

A markedly different pattern of mutation was seen in V_{λ} sequences of the sIgM loss variants from Δ XRCC2-DT40. Nearly all the sequences carried point mutations, typically with multiple point mutations per sequence (Fig. 2). A substantial shift towards point mutations was also seen in the sequences from the sIgM⁻ Δ XRCC3-DT40 cells. Thus, whereas a V_{λ} -inactivating mutation in wild-type DT40 is most likely to reflect an out-of-frame gene conversion tract, in Δ XRCC2/3 it is likely to be a point mutation (Fig. 2c). Furthermore, whereas most of the non-functional V_{λ} sequences obtained from sorted sIgM loss variants of Δ XRCC2-DT40 (53%) or Δ XRCC3-DT40 (64%) carried additional point mutations as well as the V_{λ} -inactivating mutation, such hitchhiking was only rarely observed in the non-functional V_{λ} sequences from the parental DT40 line (7%; Fig. 2d).

All of these observations point to the hypothesis that the high prevalence of sIgM-loss variants in Δ XRCC2/3-DT40 cells might simply reflect a very high frequency of spontaneous immunoglobulin V gene hypermutation in these cells. We therefore analysed the spontaneous V_{λ} mutation frequency in wild-type and Δ XRCC2/3-DT40 cells by amplifying with PCR the rearranged V_{λ} segments from total (unsorted) DT40 populations that had been expanded for 1 month after subcloning. The result shows that there is indeed a much higher spontaneous accumulation of mutations in the Δ XRCC2 and Δ XRCC3 cells than in the parental DT40 line (Fig. 3a, b). (In addition to the mutations shown in Fig. 3b, one Δ XRCC2-DT40 sequence contained a 2-basepair (bp) insertion in the leader intron, which was not obviously templated from a donor pseudogene, and one Δ XRCC3-DT40 sequence carried a single-base-pair deletion in the leader intron.) In Δ XRCC2-DT40 cells, mutations accumulate in V_{λ} at a rate of about 0.4×10^{-4} per bp per generation (given a division time of about 12 h), a value similar to

that seen in the constitutively mutating, human Burkitt lymphoma line Ramos¹⁶.

Somatic hypermutation in germinal-centre B cells in man and mouse is preferentially targeted to the rearranged immunoglobulin V_H and variable light-chain (V_L) segments. A similar situation applies to the point mutations in Δ XRCC2-DT40 cells. Thus, a significant level of apparent point mutation is also seen in the productively rearranged V_H1 gene (Fig. 3c). However, this does not reflect a general mutator phenotype, as accumulation of mutations is much lower in the light-chain constant region (C_{λ}) than in the rearranged V_{λ} and is also low in the unrearranged V_{λ} on the excluded allele where the apparent mutation rate does not rise above the background level ascribable to the PCR amplification itself (Fig. 3c; see Methods).

The distribution of mutations over the V_{λ} domain in Δ XRCC2-DT40 cells is markedly non-random. The mutations, which are predominantly single-nucleotide substitutions, show preferential accumulation at hotspots that conform to an AGY (where Y is pyrimidine) consensus on one of the two DNA strands (Fig. 3d). Mutations also show overwhelming accumulation (96%) at G/C (Fig. 3e). This G/C-biased, hotspot-focused hypermutation in Δ XRCC2-DT40 cells, although not exhibiting an obvious bias in favour of nucleotide transitions, is otherwise markedly similar to the pattern of V gene hypermutation described in cultured human Burkitt lymphoma cells as well as that occurring *in vivo* in frog, shark and MSH2-deficient mice^{17,18}. The immunoglobulin V gene hypermutation that occurs *in vivo* in man and in normal mice seems to be achieved by this hotspot-focused, G/C-biased component acting in concert with a mechanism that targets A/T. The mechanism of immunoglobulin V gene hypermutation in Δ XRCC2/3-DT40 cells is likely to be similar to that responsible for the G/C-biased phase of mutation in man and mouse. This accords well with the fact that bursal lymphoma cell lines show many parallels to the human Burkitt lymphoma lines in which immunoglobulin V gene hypermutation takes place. Both Ramos and DT40 are p53^{-/-},

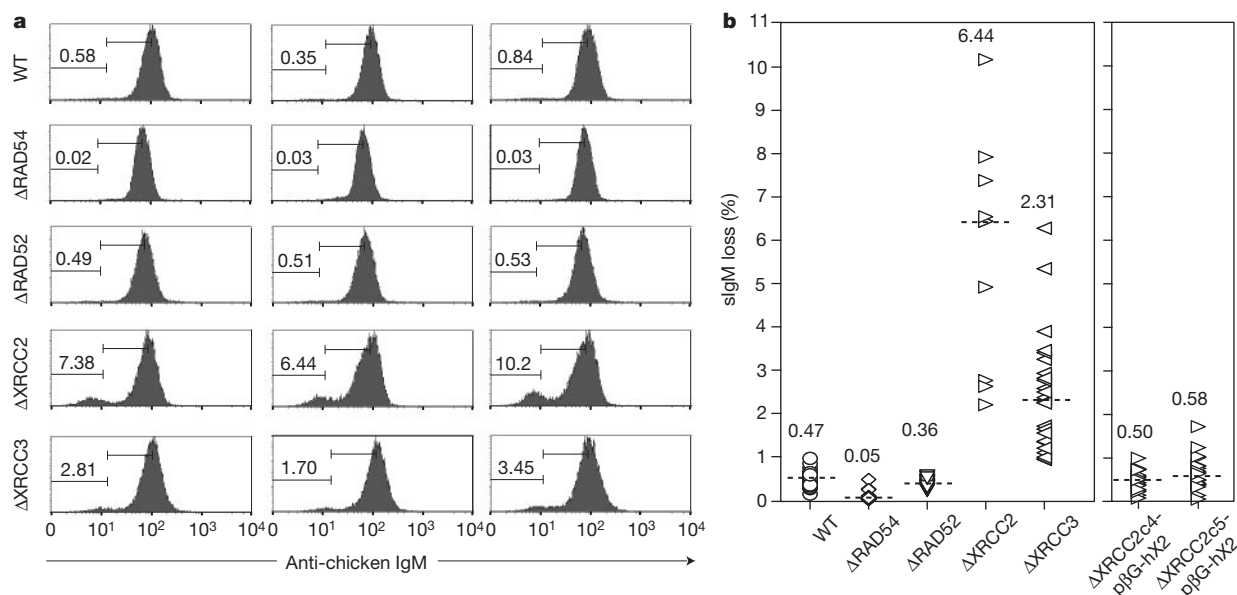


Figure 1 Generation of sIgM-loss variants by wild-type and repair-deficient DT40 cells. **a**, Representative flow cytometric analyses of the heterogeneity of sIgM expression in cultures derived by 1 month of clonal expansion of single sIgM⁺ normal (WT) or repair-deficient DT40 cells. An analysis of cultures derived from three representative sIgM⁺ precursor clones is shown for each type of repair-deficient DT40. **b**, Fluctuation analysis of the generation frequency of sIgM-loss variants. The abundance of sIgM-loss variants was determined in several parallel cultures derived from sIgM⁺ single cells after clonal

expansion (1 month); median percentages are noted above each data set and are indicated by the dashed line. The (pRG-hXRCC2) Δ XRCC2 cells are transfected with a human XRCC2 cDNA expression construct. The whole analysis was performed on multiple independent sIgM⁺ clones (with distinct, although similar, ancestral V_{λ} sequences) giving, for each repair-deficient line, average median frequencies at which sIgM-loss variants were generated after 1 month. WT, 0.4%; Δ RAD54, 0.07%; Δ RAD52, 0.4%; Δ XRCC2, 6%; and Δ XRCC3, 2%.

surface IgM⁺ B cells that express activation-induced deaminase¹⁹ (R. S. Harris and M.S.N., unpublished observations) and carry a c-MYC that has been activated by chromosomal translocation or retroviral integration. The main difference between immunoglobulin V gene mutation in human Burkitt lines and chicken DT40 cells is the higher proportion of transversion mutations observed with DT40 (Fig. 3e): this might reflect a difference in the polymerases involved in the process.

These results reveal that the DT40 chicken bursal lymphoma line, which normally exhibits a low frequency of immunoglobulin V gene diversification by gene conversion, can be induced to perform a high frequency of constitutive immunoglobulin V gene somatic mutation by ablating XRCC2 or XRCC3. This provides strong support to the earlier proposal that immunoglobulin V gene conversion and hypermutation might constitute different ways of resolving a

common DNA lesion^{20,21}. This does not mean that modulation of XRCC2/3 levels will be the only way of shifting the balance between gene conversion and somatic hypermutation. Indeed, as with a previous proposal²¹ and experiment²², we suspect a major reason that gene conversion is rarely observed in man and mouse may relate to the lack of closely located V gene donors present in accessible chromatin.

The initiating lesion for somatic hypermutation may well be a DNA break^{16,23,24}. It therefore seems significant that both XRCC2 and XRCC3 have been implicated in a recombination-dependent pathway of DNA break repair^{2,9–12}. Indeed, a similar induction of immunoglobulin V gene hypermutation in DT40 cells is achieved by ablating another gene (*RAD51B*) whose product is implicated in recombination-dependent repair of breaks²⁵, but not by ablating genes for Ku70 and DNA-PK_{cs} (the catalytic subunit of DNA-

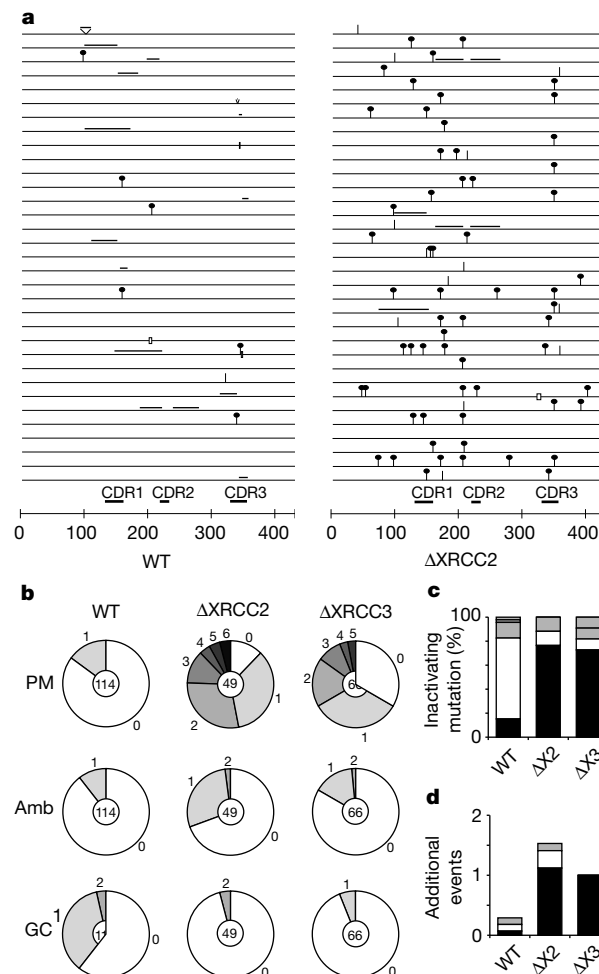


Figure 2 Analysis of V_λ sequences cloned from slgM-loss variants. **a**, Comparison of V_λ sequences from slgM-loss cells sorted from parental slgM⁺ clones of normal or XRCC2-deficient DT40 cells after clonal expansion (1 month). Each horizontal line represents the rearranged V_λ1/J_λ (427 bp) with mutations classified as described in Methods as point mutations (lollipop shape), gene conversion tracts (horizontal bar above line) or single-nucleotide substitutions, which could be a result of point mutation or gene conversion (ambiguous, vertical bar). Hollow boxes, deletions; triangles, duplications. **b**, Proportion of V_λ sequences carrying different numbers of point mutations (PM), gene conversions (GC) or mutations of ambiguous origin (Amb) among sorted slgM-loss populations. Segment sizes are proportional to the number of sequences carrying the number of mutations indicated around the periphery of the pie charts. The total number of V_λ sequences analysed is indicated in the centre of each chart with the data compiled from analysis of four subclones of wild-type DT40, two of ΔXRCC2-DT40 and three of ΔXRCC3-DT40.

Deletions, duplications and insertions are excluded from this analysis. In wild-type cells, there were additionally 6 deletions, 1 duplication and 1 insertion. There were no other events in ΔXRCC2-DT40, and a single example each of a 1-bp deletion and a 1-bp insertion in the ΔXRCC3-DT40 database. **c**, Causes of V_λ gene inactivation in wild-type, ΔXRCC2 (ΔX2) and ΔXRCC3 (ΔX3) DT40 cells expressed as a percentage of the total sequences containing an identified inactivating mutation. Missense mutation, black; gene conversion-associated frameshift, white; deletions, insertions or duplication-associated frameshift, grey. **d**, Additional mutational events associated with each inactivating mutation. Data are expressed as the mean number of additional mutations associated with each inactivating mutation with the type of additional mutation indicated as in **c**. Thus, ΔXRCC2-DT40 has a mean of 1.2 additional point mutations in addition to the index inactivating mutation, whereas wild-type DT40 has only 0.07.

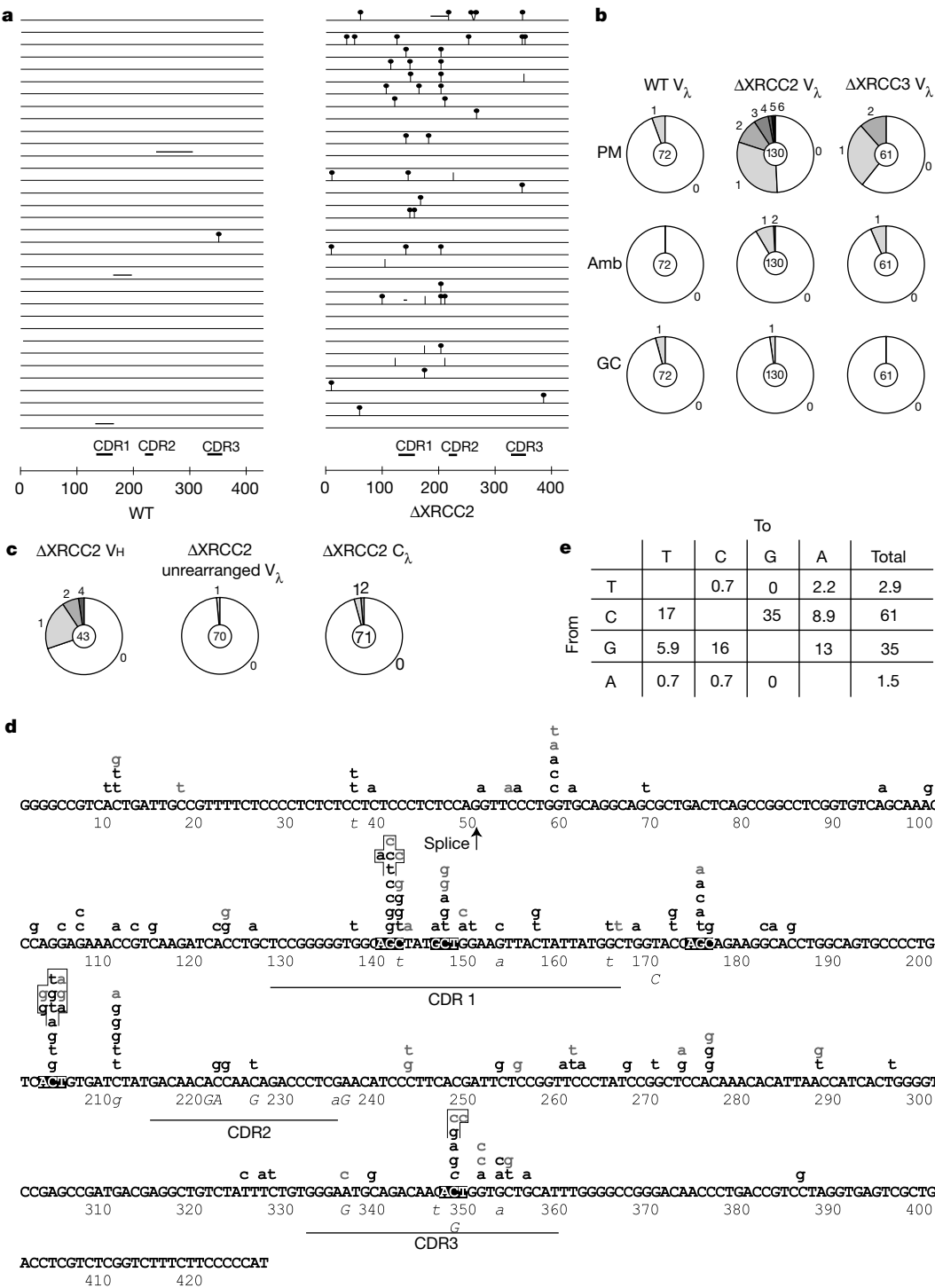


Figure 3 Analysis of immunoglobulin sequences cloned from unsorted DT40 populations after 1 month of clonal expansion. **a**, V_{λ} sequences obtained from representative wild-type and $\Delta XRCC2$ -DT40 clones; symbols as in Fig. 2. **b**, Proportion of V_{λ} sequences carrying different numbers of the various types of mutation. Data are pooled from analysis of independent clones: wild type, two clones; $\Delta XRCC2$, four clones; $\Delta XRCC3$, two clones. **c**, Mutation at other loci of $\Delta XRCC2$ -DT40. Proportion of sequences derived from $\Delta XRCC2$ -DT40 cells (after 1 month of clonal expansion) carrying mutations in the rearranged V_H (272 bp extending from CDR1 to the end of J μ), in the unarranged V_{λ} on the excluded allele (458 bp), and in the vicinity of C_{λ} (425 bp extending from the J λ - C_{λ} intron into the first 132 bp of C_{λ}). **d**, Distribution of point mutations across $V_{\lambda}1$. The consensus $V_{\lambda}1$ sequence in one of the parental slgM^+ $\Delta XRCC2$ -DT40 clones is indicated (upper case) with the first base corresponding to the 76th base pair of the leader intron. Positions at which the consensus sequences differ among the various parental slgM^+ clones are indicated (italics) below the line: differences characteristic of the

$\Delta XRCC3$ -DT40 parental clone (upper case); sporadic differences between individual parental slgM^+ subclones (lower case). Lower-case letters above the consensus indicate mutations with those from $\Delta XRCC2$ -DT40 in black and those from $\Delta XRCC3$ -DT40 in mid-grey. All mutations falling into the point mutation and ambiguous categories are included. Correction has been made for clonal expansion as described¹⁶, so each lower-case letter represents an independent mutational event. (Most of the 27 mutations thereby removed from the original database of 158 were at one of the seven principal hotspots; the correction for clonality will, if it gives rise to any distortion, lead to an underestimate of hotspot dominance.) Of the seven principal hotspots (identified by an accumulation of ≥ 5 mutations), five conform to the AGY consensus sequence on one of the two strands (black boxes). **e**, Nucleotide substitution preferences (as a percentage of the database of 131 independent events) deduced from point mutations in sequences from unselected $\Delta XRCC2$ - and $\Delta XRCC3$ -DT40. A similar pattern of preferences is evident if the $\Delta XRCC2$ / $\Delta XRCC3$ databases are analysed individually.

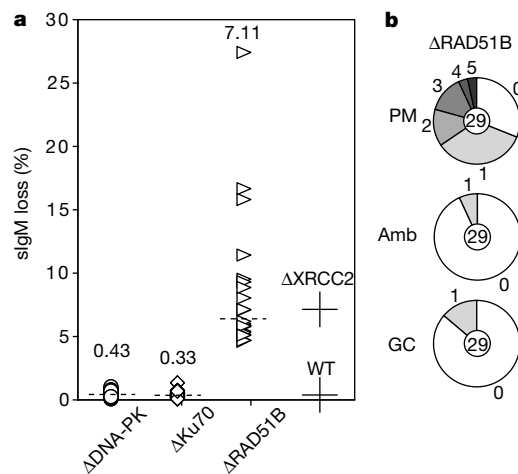


Figure 4 Analysis of slgM-loss variants in DT40 cells deficient in DNA-PK, Ku70 and RAD51B. **a**, Fluctuation analysis of the frequency of generation of slgM-loss variants after 1 month of clonal expansion. The median values obtained with wild-type and Δ XRCC2-DT40 cells are included for comparison. **b**, Proportion of V_{λ} sequences amplified from the slgM-loss variants derived from two slgM⁺ RAD51B-deficient DT40 clones carrying various types of mutation. In addition, one sequence carried a 9-bp deletion, one carried a 4-bp duplication and one carried a single-base-pair insertion.

dependent protein kinase), which are involved in non-homologous end-joining (Fig. 4).

The results, however, do not simply suggest that in the absence of XRCC2 a lesion that would normally be resolved by gene conversion is instead resolved by a process leading to somatic hypermutation. First, Δ XRCC2-DT40 cells retain the ability to perform immunoglobulin V gene conversion, albeit at a reduced level (Fig. 3b). Second, the frequency of hypermutation in Δ XRCC2-DT40 cells is about an order of magnitude greater than the frequency of gene conversion in the parental DT40 line.

We therefore propose that, in normal DT40 cells, only a small proportion of the lesions in the immunoglobulin V gene are subjected to templated repair from an upstream pseudogene, thereby leading to the observed gene conversion events. A large proportion of the lesions in the immunoglobulin V gene are subjected to a recombinational repair using the identical V gene located on the sister chromatid as a template, and which is therefore 'invisible'. This would be consistent with the observation²³ that detectable immunoglobulin V gene breaks in hypermutating mammalian B cells are restricted to the G2/S phase. In the absence of XRCC2, XRCC3 or RAD51B, we suggest that the 'invisible' sister-chromatid-dependent recombinational repair is perverted, thereby resulting in hypermutation. This hypermutation observed in the absence of XRCC2/XRCC3 could reflect either that sister-chromatid-dependent recombinational repair has become error prone, or that it has been inhibited, thereby revealing an alternative, non-templated mechanism of break resolution. Resolving which one of these two mechanisms applies will be important not only for an understanding of the mechanism of hypermutation, but also because it may provide an insight into the physiological function of RAD51 paralogues. Our identification of cell lines that perform immunoglobulin V gene hypermutation at high frequency and that constitute suitable hosts for directed gene targeting should facilitate further study of these issues. □

Methods

Cell culture, transfection and analysis

DT40 subclone CL18 and its mutants were propagated in RPMI 1640 medium

supplemented with 7% fetal calf serum, 3% chicken serum (Life Technologies), 50 μ M 2-mercaptoethanol, penicillin and streptomycin, at 37 °C in 10% CO₂. Cell density was maintained at 0.2–1.0 $\times 10^6$ ml⁻¹ by splitting the cultures daily. The generation of the DT40 derivatives carrying targeted gene disruptions has been described^{2,7,8,25,26}. Transfectants of two slgM⁺ Δ XRCC2-DT40 clones (that had given rise to slgM⁻ cells at a prevalence of 6.4 and 10.2% in the fluctuation analysis) were established by electroporation of a pSV2-*neo*-based plasmid containing the XRCC2 open reading frame (cloned from HeLa cDNA) under control of the β -globin promoter.

CL18 is an slgM⁻ subclone of DT40 and is the parental clone for the DNA repair mutants described here. Multiple slgM⁺ subclones were obtained from both wild-type and repair-deficient mutants using a Mo-Flo (Cytomation) sorter after staining with FITC-conjugated goat anti-chicken IgM (Bethyl Laboratories). Reversion of the single-base-pair frameshift in both the CDR1 region of CL18 V_{λ} and in that of its Δ XRCC2 and Δ XRCC3 derivatives (which occurs at a lower rate), is most commonly achieved by a gene conversion using the ψ V8 donor. This is the reversion found in all subclones used in this study, thereby minimizing any possible effect of consensus-sequence variation on the conversion–mutation spectrum. However, the individual Δ XRCC2 and Δ XRCC3 slgM⁺ revertants used carried a small number of nucleotide differences from each other within $V_{\lambda}1$ (see Fig. 3c), and these were taken into account when analysing mutation.

The frequency of generation of slgM-loss variants was monitored by flow cytometric analysis of cells that had been expanded for 1 month after cloning, and then stained with FITC-conjugated goat anti-chicken IgM. In each analysis, the abundance of slgM⁻ cells was determined as the percentage of live cells (as judged by propidium iodide uptake and scatter gating) whose FITC fluorescence fell at least eightfold below that of the slgM⁺ peak (see Fig. 1a).

Mutation analysis

Genomic DNA was amplified by PCR from 5,000 cell equivalents using PfuTurbo (Stratagene) polymerase and a hotstart touchdown PCR (8 cycles at 95 °C (1 min), 68–60 °C (at 1 °C per cycle) (1 min), 72 °C (1.5 min); 22 cycles at 94 °C (30 s), 60 °C (1 min), 72 °C (1.5 min)). The rearranged V_{λ} was amplified using CVLF6 (5'-CAGGAGCTCGC GGGGCGCTCACTGATTGCCG; priming in the leader- V_{λ} intron) and CVLR3 (5'-GCGCAAGCTTCCCGAGCTGCGCCCAAGTCCAAG; priming back from 3' of the light-chain joining segment (J_{λ})); the unrearranged $V_{\lambda}1$ using CVLF6 with CVLURR1 (5'-GGAATTCTCAGTGGGAGCAGGAGCAG); the rearranged $V_{\lambda}H$ using CVH1F1 (5'-CGGGAGCTCCGTCAGCGCTCTCTGTCC) with CJH1R1 (5'-GGGTACCCGGAGGA GACGATGACTTCGG) and the C_{λ} region using CJCIR1F (5'-GCAGTTCAAGAATTCC TCGGTGG; priming from within the J_{λ} – C_{λ} intron) with CCMUCLAR (5'-GGAGCCATC GATCACCAATCCAC; priming back from within C_{λ}). After purification on QIAquick spin columns (Qiagen), PCR products were cut with the appropriate restriction enzymes, cloned into pBluescriptSK and sequenced using the T3 or T7 primers and an ABI377 sequencer (Applied Biosystems). Sequence alignment²⁷ with GAP4 allowed identification of changes from the consensus sequence of each clone.

All sequence changes were assigned to one of three categories: gene conversion, point mutation or an ambiguous category. This discrimination rests on the published sequences of the V_{λ} pseudogenes that could act as donors for gene conversion. DT40 is derived from an F₁ hybrid bird with two immunoglobulin light-chain (λ) alleles, G4 and S3; the rearranged allele⁵ being G4. The database of donor sequences was taken from ref. 28, but we implemented the modifications pertaining to the immunoglobulin λ G4 allele identified by ref. 29. (The sequences/gene conversions identified in this work supported the validity of this ψ V₈ sequence database.) For each mutation the database of V_{λ} pseudogenes was searched for potential donors. If no pseudogene donor containing a string ≥ 9 bp could be found then it was categorized as an untemplated point mutation. If such a string was identified and there were further mutations that could be explained by the same donor, then all these mutations were assigned to a single gene-conversion event. If there were no further mutations then the isolated mutation could have arisen through a conversion mechanism or could have been untemplated, and was therefore categorized as ambiguous.

Analysis of known $V_{\lambda}H$ pseudogene sequences³⁰ did not indicate that any of the mutations observed in the rearranged $V_{\lambda}H$ (Fig. 3c) were due to gene conversion, suggesting that they may be due to point mutation. The mutation prevalences in these data sets are: 1.6×10^{-3} mutations per base pair for $V_{\lambda}H$; 0.03×10^{-3} for the unrearranged $V_{\lambda}1$ and 0.13×10^{-3} for C_{λ} as compared to 2.0×10^{-3} for point mutations in the rearranged $V_{\lambda}1$ in Δ XRCC2-DT40; 0.13×10^{-3} for point mutations in rearranged $V_{\lambda}1$ in wild-type DT40; and 0.04×10^{-3} for background PCR error.

Received 25 April; accepted 24 July 2001.

- Honjo, T. & Alt, F. W. *Immunoglobulin Genes* (Academic, London, 1995).
- Takata, M. et al. Chromosome instability and defective recombinational repair in knockout mutants of the five Rad51 paralogs. *Mol. Cell. Biol.* **21**, 2858–2866 (2001).
- Baba, T. W., Giroir, B. P. & Humphries, E. H. Cell lines derived from avian lymphomas exhibit two distinct phenotypes. *Virology* **144**, 139–151 (1985).
- Buerstedde, J. M. et al. Light chain gene conversion continues at high rate in an ALV-induced cell line. *EMBO J.* **9**, 921–927 (1990).
- Kim, S., Humphries, E. H., Tjoelker, L., Carlson, L. & Thompson, C. B. Ongoing diversification of the rearranged immunoglobulin light-chain gene in a bursal lymphoma cell line. *Mol. Cell. Biol.* **10**, 3224–3231 (1990).
- Sayegh, C. E., Drury, G. & Ratcliffe, M. J. Efficient antibody diversification by gene conversion *in vivo* in the absence of selection for V(D)J-encoded determinants. *EMBO J.* **18**, 6319–6328 (1999).
- Bezzubova, O., Silbergleit, A., Yamaguchi-Iwai, Y., Takeda, S. & Buerstedde, J. M. Reduced X-ray resistance and homologous recombination frequencies in a RAD54-/- mutant of the chicken DT40 cell line. *Cell* **89**, 185–193 (1997).

8. Yamaguchi-Iwai, Y. *et al.* Homologous recombination, but not DNA repair, is reduced in vertebrate cells deficient in RAD52. *Mol. Cell. Biol.* **18**, 6430–6435 (1998).
9. Liu, N. *et al.* XRCC2 and XRCC3, new human Rad51-family members, promote chromosome stability and protect against DNA cross-links and other damages. *Mol. Cell* **1**, 783–793 (1998).
10. Johnson, R. D., Liu, N. & Jasin, M. Mammalian XRCC2 promotes the repair of DNA double-strand breaks by homologous recombination. *Nature* **401**, 397–399 (1999).
11. Pierce, A. J., Johnson, R. D., Thompson, L. H. & Jasin, M. XRCC3 promotes homology-directed repair of DNA damage in mammalian cells. *Genes Dev.* **13**, 2633–2638 (1999).
12. Brenneman, M. A., Weiss, A. E., Nickoloff, J. A. & Chen, D. J. XRCC3 is required for efficient repair of chromosome breaks by homologous recombination. *Mutat. Res.* **459**, 89–97 (2000).
13. Cui, X. *et al.* The XRCC2 and XRCC3 repair genes are required for chromosome stability in mammalian cells. *Mutat. Res.* **434**, 75–88 (1999).
14. Deans, B., Griffin, C. S., Maconochie, M. & Thacker, J. Xrcc2 is required for genetic stability, embryonic neurogenesis and viability in mice. *EMBO J.* **19**, 6675–6685 (2000).
15. Griffin, C. S., Simpson, P. J., Wilson, C. R. & Thacker, J. Mammalian recombination-repair genes XRCC2 and XRCC3 promote correct chromosome segregation. *Nature Cell Biol.* **2**, 757–761 (2000).
16. Sale, J. E. & Neuberger, M. S. TdT-accessible breaks are scattered over the immunoglobulin V domain in a constitutively hypermutating B cell line. *Immunity* **9**, 859–869 (1998).
17. Rada, C., Ehrenstein, M. R., Neuberger, M. S. & Milstein, C. Hot spot focusing of somatic hypermutation in MSH2-deficient mice suggests two stages of mutational targeting. *Immunity* **9**, 135–141 (1998).
18. Diaz, M., Flajnik, M. F. & Klinman, N. Evolution and the molecular basis of somatic hypermutation of antigen receptor genes. *Phil. Trans. R. Soc. Lond. B* **356**, 67–72 (2001).
19. Muramatsu, M. *et al.* Class-switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* **102**, 553–563 (2000).
20. Maizels, N. Somatic hypermutation: how many mechanisms diversify V region sequences? *Cell* **83**, 9–12 (1995).
21. Weill, J. C. & Reynaud, C. A. Rearrangement/hypermutation/gene conversion: when, where and why? *Immunol. Today* **17**, 92–97 (1996).
22. Xu, B. & Selsing, E. Analysis of sequence transfers resembling gene conversion in a mouse antibody transgene. *Science* **265**, 1590–1593 (1994).
23. Papavasiliou, F. N. & Schatz, D. G. Cell-cycle-regulated DNA double-stranded breaks in somatic hypermutation of immunoglobulin genes. *Nature* **408**, 216–221 (2000).
24. Bross, L. *et al.* DNA double-strand breaks in immunoglobulin genes undergoing somatic hypermutation. *Immunity* **13**, 589–597 (2000).
25. Takata, M. *et al.* The Rad51B promotes homologous recombinational repair. *Mol. Cell. Biol.* **20**, 6476–6482 (2000).
26. Takata, M. *et al.* Homologous recombination and non-homologous end-joining pathways of DNA double-strand break repair have overlapping roles in the maintenance of chromosomal integrity in vertebrate cells. *EMBO J.* **17**, 5497–5508 (1998).
27. Bonfield, J. K., Smith, K. & Staden, R. A new DNA sequence assembly program. *Nucleic Acids Res.* **23**, 4992–4999 (1995).
28. Reynaud, C. A., Anquez, V., Grimal, H. & Weill, J. C. A hyperconversion mechanism generates the chicken light chain preimmune repertoire. *Cell* **48**, 379–388 (1987).
29. McCormack, W. T., Hurley, E. A. & Thompson, C. B. Germ line maintenance of the pseudogene donor pool for somatic immunoglobulin gene conversion in chickens. *Mol. Cell. Biol.* **13**, 821–830 (1993).
30. Reynaud, C. A., Dahan, A., Anquez, V. & Weill, J. C. Somatic hyperconversion diversifies the single V_H gene of the chicken with a high incidence in the D region. *Cell* **59**, 171–183 (1989).

Acknowledgements

We thank A. Johnson for cell sorting and C. Milstein for helpful discussions.

Correspondence should be addressed to J.E.S. (e-mail: jes@mrc-lmb.cam.ac.uk) or M.S.N. (e-mail: msn@mrc-lmb.cam.ac.uk).

erratum

Nodulation of legumes by members of the β -subclass of Proteobacteria

Lionel Moulin, Antonio Munive, Bernard Dreyfus & Catherine Boivin-Masson

Nature **411**, 948–950 (2001).

In Figs 1 and 3 the tree branches were very faint. The corrected figures are shown below. □

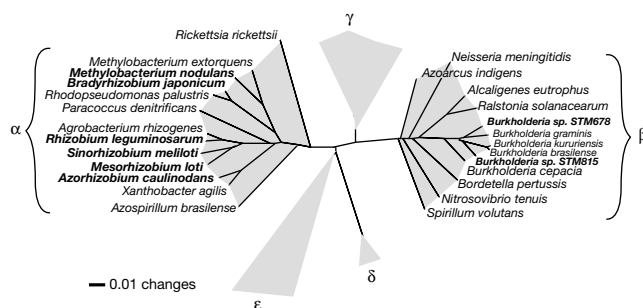


Figure 1

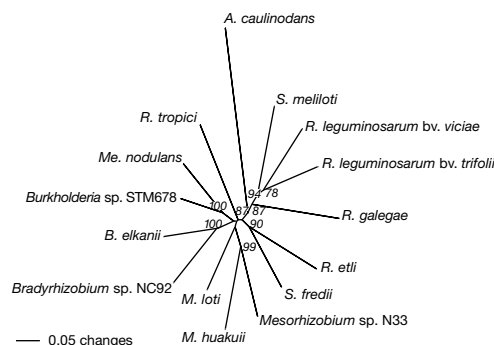


Figure 3

Going separate ways

The latest developments in chromatography.

Breeze

Waters

www.waters.com

Breeze through basic HPLC

The Breeze HPLC system consists of a binary or isocratic pump, automated or manual injector, UV or refractive-index detector and Breeze software. It offers high-performance liquid chromatography to laboratories needing basic HPLC-derived information at an affordable price.

CIS 3 S

Gerstel

www.gerstel.com

Thermal desorption system

A new cooled injection system (CIS) enables Gerstel thermal desorption systems to be used with gas chromatography systems such as the Varian 3800GC, the Saturn GC/MS or the Perkin Elmer GC. Using the Gerstel Twister, it provides complete sample recovery and lowers detection capability to the ppt and ppq ranges. The cooled injection port serves as a refocusing module for thermal desorption, providing headspace, large-volume liquid and SPME injections. The port functions in seven different modes and is optimized for capillary column GC columns with septumless head for ultra-low bleed and no sample discrimination.

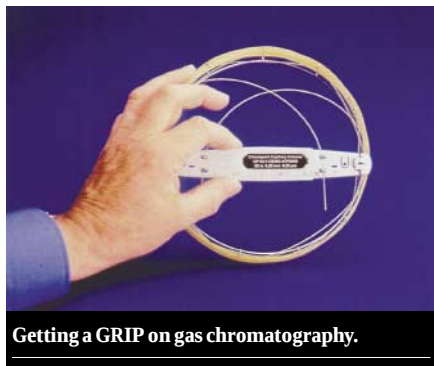
EZ-GRIP Chrompack capillary columns

Varian

www.varianinc.com

Gripping instalment

EZ-GRIP Chrompack capillary columns have a toolbar that uses a series of tabs to hold column elements during installation. These tabs simplify pre-column/retention gap and dual-column installation, ensuring secure connection of single ferrule connectors. The tabs also help to identify column inlet and outlet ends for proper installation. They can be used with single and dual gas-chromatography columns and the manufacturer claims that they can reduce installation time by as much as 90% over standard columns.



Getting a GRIP on gas chromatography.



Life's a breeze with Waters' high-performance liquid chromatography system.

Handee devices

Pierce

www.piercenet.com

Small helpings

Designed for manipulating small volumes of chromatography supports, these devices are available in three formats: spin cup/mini-spin columns, resin separators and micro-centrifuge tubes. Chromatography resin and samples are applied to the cup and a micro-centrifuge is used to wash and elute the sample. Alternatively, Luer-lok adaptors can be purchased to do washes and elutions with a syringe, using a microcentrifuge tube. The resin separators are suitable for larger volumes (250 µl–1 ml) of resin, with no need to use a microcentrifuge or a syringe. The resin separator slides into a 16 mm × 100 mm test tube containing the support and sample. The separator compresses the resin away from the sample, which is then decanted out of the tube. An instruction book can be downloaded from the company's website.

γ-RAM

IN/US Systems

www.inus.com

Radio-HPLC gamma detectors for online measurement

The γ-RAM is a research-grade, online radioactivity quantification system of ^{150}I , ^{131}I , $^{99\text{m}}\text{Tc}$ and other gamma emitters for HPLC techniques. Features include thallium-activated NaI well-type crystal for counting geometry, flexible GLP software with half-life correction,

two independent radio channels, three analogue inputs for mass detectors, four analogue outputs for chart recording and chromatography systems, external control for any HPLC system and remote computer, and detector shielding for low backgrounds. Gamma and positron cells meet all applications needs. Liquid connections are located at the front panel for convenience and control of possible leaks. The system is recommended by the manufacturer for label verification and dosing in PET laboratories.

Switchos II

Dionex

www.dionex.com

Advanced microcolumn switching

Switchos II is designed for advanced microcolumn switching and is equipped with two 10-port valves, allowing more complex sample handling. It consists of a high-precision loading pump, two 10-port low-dispersion switching valves and a four-channel solvent selection valve. The switching valves ensure dead volume-free connection of any capillary- or nano-LC column, such as the Fusica, NanoSeries, Micro- and Nano-Precolumns. The solvent selection valve permits automatic or manual selection of up to four different solvents without pump shutdown or purging. Applications include affinity extractions, isolation of phosphorylated peptides from complex protein digest, automated removal of detergents, extraction of drug compounds from biological fluids and high-throughput analysis.

Injection port septa

J G Finneran

www.jgfinneran.com

Optimized for specific analyses

This range offers septa specifically designed to accommodate different analyses in gas chromatography. Bleed and temperature optimized septa are suitable for injection-port temperatures up to 400 °C and have an injection life of up to 150 punctures. Advanced Green general-purpose septum provides low bleed conditions, an injection life of up to 100 injections and heat stability to 300 °C. Long Life Yellow 11 mm septum is suitable for prolonged and unattended operation, with heat stability to 350 °C and very low bleed at 300 °C. Low Bleed Gray 11 mm septum offers ease of sample injection and an injection life of up to 100 injections. The most economical choice is the 11 mm Low Bleed Red septum.

788IC

Metrohm

www.metrohm.co.uk

Integrated sample filtration for IC

In modern ion chromatography (IC) it is recommended that samples be filtered before injection, using membrane filters with pore sizes of 0.45 µm and lower. But the relatively high price of such disposable filters limits their use. The 788 IC filtration sample processor integrates membrane filtration as a sample preparation step directly in the IC system. The system is particularly suitable for light- to medium-loaded samples such as drinking water, surface water and waste water, digestion solutions or extracts.

1100 Series

Agilent Technologies

www.agilent.com

Scalable sample purification

The 1100 Series purification system features scalable choices that match the purification process to sample size, ranging from microgram to gram quantities. A choice of pumps with flow rates up to 100 ml/min, flow cells and a high-capacity fraction collector optimize each analysis. Fraction collection is based on time, peak and mass, allowing users to collect and automatically reanalyse fraction. Both purification and purity checks are completed in one automated system.

Application note CD

Jones Chromatography

www.joneschrom.com

HPLC in the round

This CD contains 88 new applications for Genesis HPLC columns. It includes general applications and application notes pertaining specifically to pharmaceutical, environmental, biomedical and petrochemical work, as well as applications in food and beverage, quality control, vitamins, and peptides and proteins. The CD is free and can be ordered online at the company's website.



A septum to suit any occasion, from J G Finneran.

542 Peltier rack

Gilson

www.gilson.com

Large-capacity sample cooling and heating

This rack cools and heats samples from 10 °C to 40 °C and is designed for integration into the NEBULA Series 215 line of instruments. It holds ten 96-well shallow-well microplates (960 samples) in a compact space. An optional 242 Peltier rack can be added to provide cooling and heating for up to 1,152 samples. The racks are especially useful for long runs and unattended overnight assays. Their cooling capabilities reduce sample evaporation, making them suitable for smaller volume sample preparation for HPLC or LC/MS.

Chrysalis

Matheson Tri-Gas

www.mathesontrigas.com

Hydrogen generators for fast GC

These gas generators produce hydrogen at delivery pressures up to 200 psi for use as a carrier gas in GC or a fuel gas with FID, FPD and other detectors. The hydrogen produced has less than 1 ppm water and oxygen, and under 1 ppb hydrocarbons. Two models are available, both only 9 inches wide. The 300 cc model can supply up to five GCs simultaneously; the 600 cc model can supply up to 10 GCs. Operation and settings are controlled using a front-mounted LED panel. The generators use a maintenance-free palladium purifier instead of conventional desiccant cartridges to remove water and impurities. Deionized water is the only liquid necessary for operation, added through a built-in funnel. For applications requiring continuous use, an automatic filler option is available.

LumaPlate-96

Packard Instrument

www.packardinstrument.com

Bar-coded microplates

LumaPlate-96 and DeepWell LumaPlate-96 are scintillator-coated microplates constructed from polystyrene, which provides greater chemical resistance to acetonitrile for counting HPLC fractions. The 3H background count

rate is lowered by about a factor of two with the LumaPlate-96, and by about 30–40% with the DeepWell LumaPlate-96. Both microplates come with pre-applied bar code labels that can be read by the TopCount NXT microplate scintillation and luminescence counter. The standard bar code symbologies are code 128, code 39 and Interleaf 2 of 5. Other symbologies are available by special order.

High-performance chromatography columns

Omnifit

www.omnifit.co.uk

Smooth operators

These chromatography columns are precision-bored from borosilicate glass and have fully inert wetted parts, resulting in minimal dead volume, elimination of sample contamination problems and smooth solvent flow. The end-pieces are fully adjustable, allowing bed length to be varied, and the PTFE primary seals have a choice of viton, silicone or chemraz backup seals. The frit design allows rapid replacement in the event of a blockage. The columns are available in diameters of 5, 10, 16 and 27 mm and in lengths between 50 mm and one metre. Accessories include column injectors, support rods and filling funnels. Applications include bioseparations and other analytical and semi-preparative laboratory work as well as general research and production.

These notes are compiled in the Nature office from information provided by the manufacturers.

ADVERTISEMENT

Polyclonal Antibodies & Custom Peptides

Please contact us about complete services for your custom peptide and polyclonal antibody needs.

ResGen[™]
Invitrogen Corporation

U.S. or Canada 800-533-4363
FAX 256-536-9016

Check out our web site at
www.resgen.com